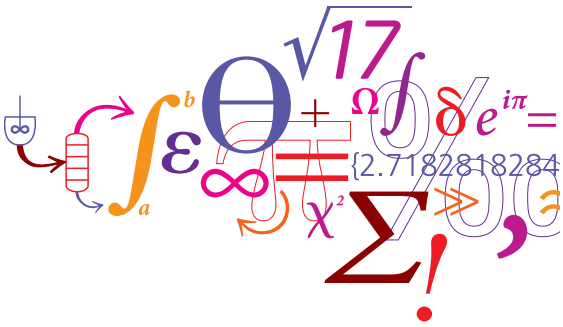
Molecular Modeling of Enzyme Dynamics Towards Understanding Solvent Effects

Wedberg, Nils Hejle Rasmus Ingemar

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Molecular Modeling of Enzyme Dynamics Towards Understanding Solvent Effects



Rasmus Wedberg
Ph.D. Thesis
February 2011

DTU Chemical Engineering
Department of Chemical and Biochemical Engineering

Molecular Modeling of Enzyme Dynamics Towards Understanding Solvent Effects

Ph.D. thesis

Rasmus Wedberg

February, 2011

Computer Aided Process Engineering Center
Department of Chemical and Biochemical Engineering
Technical University of Denmark

Address: Computer Aided Process Engineering Center
Department of Chemical and Biochemical Engineering
Technical University of Denmark
Building 229
DK-2800 Kgs. Lyngby
Denmark
Phone: +45 4525 2800
Fax: +45 4588 4588
Web: www.capec.kt.dtu.dk

Print: J&R Frydenberg A/S
København
June 2011

ISBN: 978-87-92481-45-0

Preface

This thesis is submitted as a partial fulfillment of the requirements of the PhD degree at the Technical University of Denmark ('Danmarks Tekniske Universitet'). The work has been carried at the Computer-aided process-product engineering center (CAPEC) at the Department of Chemical and Biochemical Engineering ('Institut for Kemiteknik') from November 2007 to October 2010 under the supervision of Associate Professor Jens Abildskov from the same department, and Associate Professor Günther H. Peters from the Department of Chemistry ('Institut for Kemi'). I am grateful for their exceptional presence and in particular for their willingness to devote a great deal of their time to scientific discussions, which occasionally got quite lengthy.

During two months, October and November 2008, I visited the Department of Chemical Engineering of the University of Virginia, USA, and worked under the supervision of legendary Professor John P. O'Connell. I am grateful for the ideas and encouragement he has contributed with.

I am also grateful towards IP Bioproduction, European Union Sixth Framework Programme, which has provided financial support for the project. The Danish Center of Scientific Computing is acknowledged for providing access to the Horseshoe cluster, which I have used extensively in my work.

I would like to thank every coworker and student at CAPEC for their support and for the unique and persistent social environment which I have greatly enjoyed.

Very thankful am I also towards my family and my friends from outside CAPEC for always being supportive and taking interest in the rather specific things I have worked with during these three years.

Rasmus Wedberg
Kgs. Lyngby, Denmark, February 2011

Abstract

This thesis describes the development of a molecular simulation methodology to study properties of enzymes in non-aqueous media at fixed thermodynamic water activities. The methodology is applied in a molecular dynamics study of the industrially important enzyme *Candida antarctica* lipase B (CALB) in water and organic solvents. The effects of solvent on structural and dynamical enzyme properties are studied, and special attention is given to how enzyme properties in organic solvents are affected by the hydration level, which is shown to be related to the water activity.

In experimental studies of enzyme kinetics in non-aqueous media, it has been a fruitful approach to fix the enzyme hydration level by controlling the water activity of the medium. In this work, a protocol is therefore developed for determining the water activity in non-aqueous protein simulations. The method relies on determining the concentration of water in a region of the simulation box far from the protein surface. In order to evaluate the corresponding activity, a previously developed methodology based on fluctuation solution theory is employed to compute the excess Gibbs energy of the water/organic solvent mixture. This requires that separate simulations of this mixture are carried out at different compositions, and that the total correlation function integrals, i.e. spatial integrals of the pair radial distribution functions (RDFs), are evaluated.

A main challenge is that the total correlation function integrals do not converge within the system size of the simulation box generally used in simulation. Therefore, a method is developed for extending the RDFs to arbitrary distances so that the integrals can be evaluated. The method, which was first used in the classical study of the Lennard-Jones fluid by Verlet (Verlet (1968), *Phys. Rev.*, **165**, 201–214), is here extended for application to simulations of molecular fluid mixtures. It extends the RDFs by enforcing that the corresponding direct correlation functions follow a certain approximation at long distances. This approximation is here derived in terms of statistical mechanical fluid theory. An extensive set of numerical tests are carried out for validating the method, and it is found that thermodynamic properties of good accuracy are obtained from the integrals of the extended RDFs. The method is also shown to be at least as good as existing methods for correlation function integration, while for small systems, it seems to be even better.

The method is applied to compute the excess Gibbs energy of the mixtures of water and organic solvents used in the simulations of CALB. This allows to determine the water activity of the simulated systems and thus to compare protein properties in different organic solvents at fixed water activities. The study bridges therefore the previously used simulation approach where properties were compared at similar hydration levels (Yang *et al* (2004), *Biophys. J.*, **87**, 812–821); Micaêlo and Soares (2007), *FEBS J.*, **274**, 2424–2436; Trodler and Pleiss (2008), *BMC Struct. Biol.*, **8**) and the approach to fix the water activity which often is used in experimental studies.

The water activity is shown to have a profound effect on the structure and dy-

namics of CALB. Conformational flexibility, for instance, increases with increasing hydration in acetone, t-butanol, methyl t-butyl ether and hexane, but not in methanol. A consequence of this is that hydration needs to be carefully considered in simulation studies of proteins in organic media. The organic solvent is also shown to affect structure and dynamics of CALB. The effects on flexibility can partially be attributed to the mobility of the hydration water, as proposed in a previous study (Trodler and Pleiss (2008), *BMC Struct. Biol.*, **8**). The present results indicate that flexibility may also be affected by adsorption of organic solvent molecules to the enzyme surface. This seems in particular to be the case in t-butanol in which the lowest flexibility of CALB is observed.

Future applications of the methodology may lead to an improved understanding of enzyme properties in non-aqueous media, which may have significant impact on the development of rational strategies for solvent selection in biocatalysis.

Resumé

Denne afhandling beskriver udviklingen af et molekylsimuleringsværktøj for at studere egenskaber hos enzymer i organiske medier ved konstant termodynamisk vandaktivitet. Metoden bruges i et molekyldynamisk studie af enzymet Candida antarctica lipase B (CALB) i vand og organiske opløsningsmidler. Opløsningmidlets indvirkning på enzymets struktur og dynamik studeres med fokus på hvorledes disse egenskaber påvirkes af enzymets hydratiseringsgrad, som er relateret til mediets vandaktivitet. En almindelig fremgangsmåde i eksperimentelle studier af enzymkinetik i organiske medier er atfastholde en givet hydratisering ved at regulere vandaktiviteten. Det er derfor ønskværdigt, at kunne studere vandaktivitetens indvirkning på proteiners egenskaber, også i molekylsimulering. For at bestemme vandaktiviteten, bliver først den lokale vandkoncentration, i det simulerede system langt fra proteinets overflade, bestemt. For at fastholde den tilsvarende aktivitet, bruges en fremgangsmåde baseret på *Fluctuation Solution Theory*, til at beregne overskuds-Gibbsenergien for den tilsvarende blandning af vand og organisk opløsningmiddel. Med denne fremgangsmåde simuleres væskeblandningen ved forskellige sammensætninger, for at udregne *Kirkwood-Buff* (KB)-integralerne, som er rumintegraler af molekylære parfordelingsfunktioner. Den centrale udfordring er, at disse integraler sjældent konvergerer indenfor den rækkevidde, som almindeligvis er tilgængelig i molekylsimulering. En metode er derfor blevet udviklet, til at ekstrapolere parfordelingsfunktionerne til vilkårlig rækkevidde. Metoden er baseret på en fremgangsmåde oprindeligt udviklet af Verlet (Verlet (1968), *Phys. Rev.*, **165**, 201–214) for at studere Lennard-Jones-væsker. Den er her videreudviklet, til at kunne anvendes til molekylære væskeblandinger. Metoden tilnærmer den direkte korrelationsfunktion, og er udledt fra fundamental statistisk mekanisk væsketeori. Metoden er valideret ved en række eksempler, der demonstrerer, at de beregnede KB-integraler giver termodynamiske egenskaber med stor nøjagtighed. Det er ligeledes demonstreret, at metoden er mindst lige så nøjagtig, som eksisterende metoder til beregning af KB-integraler. Metoden bruges for at beregne overskuds-Gibbsenergien for de blandninger af vand og organiske opløsningmidler, der bruges i simulering af CALB. Dette muliggør, at vandaktiviteten i de simulerede systemer kan bestemmes, og at proteinets egenskaber i forskellige organiske opløsningsmidler er sammenlignelige ved ens vandaktiviteter. Studiet viser, at CALBs struktur og dynamik påvirkes af vandaktiviteten. For eksempel øger fleksibiliteten med stigende vandaktivitet i de organiske opløsningmiddler acetone, t-butanol, metyl t-butylæter og hexan. Dette observeres dog ikke i methanol. Fleksibiliteten påvirkes også af det organiske opløsningsmiddel, hvilket delvis kan tilskrives bevægeligheden af vandmolekylerne på proteinets overflade. Den fremgangsmåde, der ved molekylsimulering er udviklet i dette arbejde, kan bruges til at opnå bedre forståelse af proteiners egenskaber i organiske opløsningmidler. Sådan et kendskab kan bruges til udvikling af systematiske strategier for design af reaktionsmedier til biokatalytiske reaktioner.

Contents

1

Introduction

Biocatalysis in non-aqueous media has been a vivid field of research since the 1980's, due to its feasibility, as well as its many advantages demonstrated by Zaks and Klibanov (1984, 1985, 1988b); Dordick *et al.* (1986); Kazandjian *et al.* (1986), among others. In contrast to conventional chemical catalysts, enzymes offer a high degree of selectivity, and do not require exceptionally high temperatures to be functional. They are for this reason often preferable as catalysts for synthesis of organic chemicals with specific molecular structures. Non-aqueous media, such as organic solvents, on the other hand offer several advantages over aqueous media from the process perspective. For instance, substrates and products for organic synthesis are usually more soluble in organic solvents, undesired side-reactions with water can be avoided and product recovery can be facilitated. Due to the shift in chemical equilibrium, organic solvents can furthermore reverse the biological function of hydrolytic enzymes, such as esterases, lipases and proteases. In aqueous solution, these enzymes are catalysts for hydrolysis while they in organic media catalyze the opposite reaction, e.g. esterification in the case of esterases. Dordick (1992) summarized potential advantages of using enzymes in organic solvents as

1. Increased solubility of non-polar substrates
2. Shifting thermodynamic equilibria
3. Suppression of side reactions involving water, e.g. hydrolysis
4. Alteration in substrate- and enantioselectivity
5. Often no need to immobilize the enzyme
6. Enzyme recovery by simple filtration
7. Easier product recovery from low boiling, high vapor pressure solvents
8. Enhanced thermo-stability
9. Elimination of microbial contamination
10. Potential for enzymes to be used directly within a new or existing chemical process

Two important contemporary applications are the production of biodiesel (Akoh *et al.*, 2007; Fjerbaek *et al.*, 2009), and production of monoacylglycerols (MAGs) (Berger and Schneider, 1992; Bellot *et al.*, 2001; Rendón *et al.*, 2001; Kaewthong and H-Kittikun, 2004; Damstrup *et al.*, 2005, 2006). The former involves transesterification of esters present in rapeseed oil with a small alcohol such as methanol or ethanol. The latter involves as well transesterification of vegetable oils and glycerol. These reactions are frequently run in organic solvent since such solvents better solubilize the substrates, among other reasons. Further applications of non-aqueous

enzymology have been reviewed by Carrea and Riva (2000).

The medium does however not only affect properties on the process scale - like those mentioned above - but also properties that are tightly linked to the molecular nature of the enzyme itself. The enzymatic activity, i.e. the kinetic rate of a catalyzed reaction is typically several orders of magnitude lower in organic media, as compared to in water, but shows also a great variation with different organic solvents (Klibanov, 1997). The medium also affects the enzyme specificity, i.e. the preference for one substrate over another, or preference for catalyzing one reaction over another (Zaks and Klibanov, 1985; Carrea *et al.*, 1995). This has important engineering implications, since specificity usually is a key property in biocatalysis. Finally, the solvent affects enzyme stability, i.e. the time the enzyme retains its catalytic activity. In some organic solvents, dramatically higher protein stability has been observed at high temperature, as compared to in water (Zaks and Klibanov, 1984).

The task of selecting an organic solvent or designing a solvent mixture in order to improve or optimize a biocatalytic process is often referred to as *medium engineering*, and sometimes described as an alternative to *protein engineering* (Carrea and Riva, 2000), in which the enzyme is improved through mutations of its primary structure. Generally, medium engineering appears simpler than protein engineering, while the latter, depending on the particular application, might provide more dramatic improvements, e.g. if the active site pocket is redesigned to accommodate larger substrates (Magnusson *et al.*, 2005). Protein engineering targets exclusively molecular-scale properties, while medium engineering mainly targets process-scale properties, many of which are enzyme-independent. It is furthermore fully possible to combine the two approaches (Wangikar *et al.*, 1993).

Hundreds of organic solvents are in use in today's industry, and in addition, these can be mixed in order to form solvents with different properties. This hints the huge potential of medium engineering, and that careful selection of solvent for a given biocatalytic process can be beneficial. However, it also implies that proper selection requires significant efforts, due to the vast number of solvent candidates to consider. Despite recent progress in modeling of complex biochemical systems, biocatalytic reaction media are mostly selected based on extrapolation of previous experience or trial and error. For instance, Su and Wei (2008) reported a medium engineering study for lipase-catalyzed biodiesel production in which a large number of experimental conversion measurements were carried out. They considered 11 pure solvents and 9 solvent mixtures, each at 10 different compositions, in order to find the best medium; a 75/25 mixture of t-pentanol and isooctane. Damstrup *et al.* (2005) approached solvent selection for MAG production in a similar fashion. They carried out experiments for 13 pure solvents, and measured the product yields. The highest yield was obtained with t-butanol as solvent.

The field of non-aqueous biocatalysis could benefit from the development of systematic, model-based approaches for screening of solvents for biocatalytic processes. Such developments are however presently limited to process parameters and solvent properties, and do not account for that the solvent may affect the behavior of the enzyme. Although such considerations are useful, as will be described in Section 2.1, it is desirable to also include enzyme properties such as activity, specificity and stability. How enzyme-solvent interactions affect such properties is however not yet

fully understood.

Zaks and Klibanov (1988a) demonstrated that in addition to the organic solvent, enzyme hydration has a significant impact on catalytic properties. In particular, enzymes are typically inactivated upon complete dehydration. This has been ascribed to that a layer of water molecules needs to be present at the enzyme surface in order for catalytic activity to be retained in organic media. It has for instance been proposed that this hydration layer is important since it acts as a lubricant providing the protein with conformational flexibility necessary for catalysis (Broos *et al.*, 1995). An alternative hypothesis is that the hydration layer ensures that the active site is hydrated (Yang *et al.*, 2004). Organic solvents differ in their ability to mix with water. Consequently, Zaks and Klibanov (1988a) reported that polar solvents, which mix well with water, were more prone to strip water from the enzyme surface, as compared to non-polar solvents. Thus, for a certain concentration of water in the reaction medium, the enzymatic activity was lowest in polar solvents, while high in non-polar ones. In a study by Laane *et al.* (1987) in which hydration was not considered explicitly, it was similarly found that the catalytic activity correlates with solvent hydrophobicity, reported as the octanol/water partition coefficient $\log P$.

Several authors have argued that hydration effects more appropriately should be studied in terms of thermodynamic water activities, rather than concentrations (Halling, 1989, 1990b; Valivety *et al.*, 1992b,a; Halling, 1994; Bell *et al.*, 1997). The idea is that the amount of water in the hydration layer depends on the medium water activity and to a lesser extent on the organic solvent. By controlling the water activity of the reaction medium, one can control the hydration of the enzyme and thus ensure that the catalytic activity is retained, even in polar solvents (Halling, 1990b). Valivety *et al.* (1992b) measured the esterification activity of *Mucor miehei* lipase as a function of water activity in various organic solvents of low polarity. A bell-shaped dependence was found with maximum enzymatic activity obtained at a water activity of 0.55. More interestingly, the enzymatic activity was rather insensitive to the solvent, as long as the water activity was kept at a fixed value. These results suggest that the solvent affects catalytic properties indirectly, by interacting with the hydration layer. A later study by Bell *et al.* (1997) investigating the effects of different water activities in polar solvents did however show that the solvent had a significant impact on the catalytic activity, even though the water activity was kept fixed. The authors reasoned that direct interactions between enzyme and solvent were responsible for this.

Molecular modeling is a useful complement to experimental studies, when one seeks to understand phenomena taking place on the molecular scale. The approach is well-suited for studying the interactions between protein, organic solvent and hydration layer. Several computational studies have already investigated the effect of non-aqueous solvent and hydration level on protein properties such as structure and dynamics, e.g. Soares *et al.* (2003); Yang *et al.* (2004); Micaêlo and Soares (2007); Trodler and Pleiss (2008); Díaz-Vergara and Piñeiro (2008) (see further Section 2.2). Simulations have also confirmed the tendency for polar solvents to reduce the hydration layer (Yang *et al.*, 2004; Micaêlo and Soares, 2007; Trodler and Pleiss, 2008; Cruz *et al.*, 2009). In order to better understand the solvent effects on protein structure and dynamics it would be valuable to be able to distinguish effects arising directly from interaction between protein and organic solvent, from effects arising

indirectly from the tendency of the solvent to preserve or reduce the hydration layer. One strategy to accomplish this is to compare protein properties obtained from simulations carried out in different solvents, at similar water activities. It seems however that only one protein simulation study has analyzed the water activity of the medium. In this study, which was performed by Branco *et al.* (2009), the medium was however not an organic solvent, but a gaseous mixture of argon and water, representing the carrier gas of a solid/gas reactor.

The objective of this PhD work is to develop a protocol for studying how structure and dynamics of proteins in organic solvents depend on the water activity, by molecular dynamics (MD) simulation. The purpose is also to apply this protocol and establish the effects of solvent choice and water activity on the industrially significant enzyme *Candida antarctica* lipase B (CALB). If accomplished, this will allow for more clear distinction between effects of organic solvent and hydration, as described above. It may also facilitate future attempts to establish correlations between molecular scale properties obtained from MD simulations, and catalytic properties measured experimentally at controlled water activity.

The thesis is organized in eight chapters. *Chapter 1* is the introduction.

- *Chapter 2* discusses various model-based approaches to study enzymes in non-aqueous media and predict properties of biocatalytic systems. A brief outline of the computer-aided molecular design method and an overview of enzyme-independent modeling approaches for solvent selection in biocatalysis are given. The focus is then shifted towards molecular modeling approaches in which the enzyme is modeled explicitly. Molecular modeling approaches including quantum mechanical computations and classical molecular simulations are briefly outlined, with particular emphasis on their role in protein science. An overview of previous molecular modeling studies of proteins in non-aqueous media is given. The importance of controlling the hydration level or water activity in such studies, which is the main objective of this thesis, is discussed.

- *Chapter 3* describes an MD study of CALB in organic media with the aim to investigate the effects of solvent and hydration level on the structure and dynamics of the protein. The main conclusion of this study is that protein properties in different solvents depend on enzyme hydration.

- *Chapter 4* discusses possible approaches for controlling or measuring enzyme hydration in simulation in terms of the thermodynamic water activity, a_{w}. In particular, a computational methodology based on fluctuation solution theory (FST) is described. This methodology is applied to analyze protein simulations in Chapter 7.

- *Chapter 5* describes specific improvements to the FST methodology, which were found to be necessary for its application to aqueous/organic fluid mixtures. The methodology relies on relating thermodynamic properties to Kirkwood-Buff integrals which are calculated from simulations. The developments presented in Chapter 5 comprise the employment of approximations derived from statistical mechanics and models for molecular interactions, to improve the accuracy of these integrals. The theoretical basis of the new method is explained.

- In *Chapter 6*, an extensive set of tests are carried out in order to validate the developments of the previous chapter. Simulations of pure and mixed fluids of molecules of varying complexity are considered. Thermodynamic properties are calculated and validated.

- A more extensive MD study of CALB is described in *Chapter 7*, in which the enzyme is simulated in pure water and five organic solvents under varying hydration conditions. The FST methodology developed in Chapters 4-5 is applied to the simulations in order to measure protein hydration in terms of a_{w}. The dependence of structural and dynamical protein properties on solvent and hydration level is discussed.

- Finally, *Chapter 8* summarizes the main conclusions of the work and gives suggestions for future directions.

The results that Chapters 5 and 6 are based on have in part been reported previously, in the journal articles

- Wedberg R., O'Connell J. P., Peters G. H., and Abildskov J. (2010). Accurate Kirkwood-Buff Integrals from Molecular Simulations. *Mol. Simul.*, **36**, 1243–1252.

- Wedberg R., O'Connell J. P., Peters G. H., and Abildskov J. (2010). Total and Direct Correlation Function Integrals from Molecular Simulation of Binary Systems. *Fluid Phase Equilib.*, (in press: DOI:10.1016/j.fluid.2010.10.004).

Other results obtained during this PhD-work have been reported in the journal article

- Wedberg R., Peters G. H., and Abildskov J. (2008). Total Correlation Function Integrals and Isothermal Compressibilities from Molecular Simulations. *Fluid Phase Equilib.*, **273**, 1–10.

2

Background

This chapter provides an overview of contemporary model-based approaches for understanding and/or predicting how the behavior of non-aqueous biocatalytic processes depends on the solvent. These approaches fall into two broad categories, namely those that do not involve the enzyme itself, and those that do. A biocatalytic process is a complex system consisting of reactants, products, solvent and biocatalyst. Much of the information which is crucial to consider when selecting solvent, involves only the three former. Such information can be obtained or estimated with relative ease, since several well-established databases and property models are available. Some examples of model-based approaches that might be useful for solvent selection are described in Section 2.1. What distinguishes biocatalysis from conventional chemical processes is the nature of the catalyst. Enzymes have very complex molecular structure exhibit as well complex dynamics (Fersht, 1999). Although interactions with the solvent are known to impact protein function, the mechanism is still not understood in detail. Rationalization of the effect of different solvents is therefore difficult. Molecular modeling approaches are promising candidates to help gaining a better understanding of solvent effects on protein dynamics. Selected molecular modeling methods are described in Section 2.2, and an overview of previous applications to non-aqueous enzymology is given.

2.1 Enzyme-Independent Modeling

This section gives an overview of contemporary modeling approaches that seeks to predict parameters for non-aqueous biocatalytic systems independently of the biocatalyst itself. The computer-aided molecular design (CAMD) method, which can be used for screening solvents for certain target properties, is briefly described. A selection of previous approaches employed to model such target properties are then described.

2.1.1 Computer-Aided Molecular Design

A few hundred solvents are in use in industry today. In addition, mixtures of these can form solvents with new properties. There are thus plenty of possibilities for optimizing a biocatalytic process through medium engineering. However, solvent selection based on experimental testing ("trial-and-error") is limited due to the requirement of time and resources. CAMD is a collective name for a set of algorithms developed to assist the selection of a molecular compound in either process or product design. The principles of these methods applied to solvent selection have been described by Achenie *et al.* (2003).

In order to formalize the design problem, a search space is defined in terms of a set of chemical groups $\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_N$ of which the candidate molecules may consist of. The structure of a molecule is represented by an integer vector $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$, where n_i specifies the number of groups of type G_i that the molecule contains. One then seeks appropriate molecular structures that satisfy a specific set of equality and inequality constraints of the type

$$\begin{aligned} \boldsymbol{h}(\boldsymbol{n}) &= 0 \\ \boldsymbol{g}(\boldsymbol{n}) &\leq 0 \end{aligned} \tag{2.1}$$

These constraints are specified in terms of certain physico-chemical "target" properties which are relevant for the particular function of the sought molecule. This commonly includes important thermodynamic properties, such as boiling point, vapor pressure, solubility of a given compound, etc. The target properties may also include ones that are important from the perspective of safety, economics or sustainability, e.g. toxicity, flammability and cost. A set of structural constraints must as well be included since any combination of groups $\boldsymbol{n}$ does not correspond to a feasible molecular structure. These constraints relate e.g. the total number of groups in the molecule to the number of available bonds (Odele and Macchietto, 1993).

The problem formulation may involve a specific performance criterion, which is a function of the molecular structure which one desires to be as large as possible. For instance, Odele and Macchietto (1993) who considered solvent selection for gas adsorption, used as performance criterion the solvent selectivity for adsorption of hydrogen sulphide relative to carbon dioxide. If a performance criterion is included in the formulation, the design task is an optimization problem of the mixed-integer nonlinear programming type, and is solved by iterative methods (Odele and Macchietto, 1993). If no such criterion is included, the task is to identify all molecular structures that satisfy the constraint, which sometimes is achieved by enumeration of all possible structures (Abildskov *et al.*, 2010b).

It is essential that the target properties can be predicted from the group composition $\boldsymbol{n}$. For this purpose, group contribution methods are commonly employed. Such methods are established for a range of thermodynamic properties (Poling *et al.*, 2007). In practical use of CAMD methods, the primary interest is not to determine "the" optimal solvent, but rather to identify a limited number of feasible solvent candidates. For this reason, the employed property models need not have high accuracy but need merely to reproduce the gross features of experimental data. Computational efficiency is however important since the search algorithm needs to test a large number of structures. Having reduced the search space to a handful of solvent candidates, the final selection is made on the basis of experimental testing, with or without guidance of more sophisticated modeling considerations.

A possible application of CAMD to solvent selection for biocatalysis in non-aqueous media was described by Abildskov *et al.* (2010b). This study focused on designing the solvent for two transesterification reactions catalyzed by *Candida antarctica* lipase B (CALB), namely those of octanol with vinyl laurate and inulin (a polysaccharide) with vinyl laurate. The property constraints were defined in terms of solubility of the substrates, the solvent-induced shift of chemical equilibrium composition, toxicity and boiling point. A further constraint was related to the miscibility with water. The reason behind this was the desire to choose a solvent that mixes poorly

with water, since interactions between the solvent and the hydration layer around the enzyme are undesirable. No performance criterion was employed in the study.

2.1.2 Substrate Solubility

Solubility refers to how high the concentration of a particular solute at equilibrium can be in a given solvent, before the solution splits into two or more distinct phases. Since the substrates used for biocatalytic reactions usually are liquids, prediction of their solubilities is accomplished by liquid-liquid equilibrium calculations. Elementary solution thermodynamics dictates that if two liquid phases α and β are in equilibrium, the fugacities of each chemical species is equal in the two phases. Expressed in terms of mole fractions and activity coefficients, this defines a system of equations for the composition of phases α and β, according to (Smith *et al.*, 2005)

$$x_{\alpha,i}\gamma_{\alpha,i} = x_{\beta,i}\gamma_{\beta,i}, \forall i \tag{2.2}$$

where $x_{\alpha,i}$ and $\gamma_{\alpha,i}$ denote respectively the mole fraction and activity coefficient of species i in phase α. Equation (2.2) holds for all chemical species, including substrates and products as well as the solvent. It should also be noted that the $\gamma_{\alpha,i}$ are (usually non-linear) functions of the composition of phase α. Equation (2.2) is however inconvenient to use for solvent selection, since the solubility of a given substrate in the reaction mixture generally depends on the extent to which other substrates and products are present. The system of equations can be simplified by considering only the solvent and one substrate at a time, neglecting the influence of the other substrates and products. This is reasonable, since the concentrations of substrates and products typically are around 1% or less for a biocatalytic process, which is rather low. For a binary mixture of a substrate and a solvent denoted respectively by 1 and 2, Equation (2.2) takes the form

$$\begin{cases} x_{\alpha,1}\gamma_{\alpha,1} &= x_{\beta,1}\gamma_{\beta,1} \\ (1-x_{\alpha,1})\,\gamma_{\alpha,2} &= (1-x_{\beta,1})\,\gamma_{\beta,2} \end{cases} \tag{2.3}$$

Solution of this system yields $x_{\alpha,1}$, which is the substrate solubility in the given solvent. Here, α is chosen to denote the solvent-rich phase. Solving Equation (2.3) requires that models for the activity coefficients are employed. The equation can however be further simplified for the purpose of CAMD for screening of solvents. If one assumes that the substrate is dilute in phase α, and the solvent is dilute in phase β, the first identity of Equation (2.3) can be rewritten as

$$\frac{x_{\alpha,1}}{x_{\beta,1}} = \frac{\gamma_{\beta,1}}{\gamma_{\alpha,1}} \approx (\gamma^{\infty}_{\alpha,1})^{-1} \tag{2.4}$$

where $\gamma^{\infty}_{\alpha,1}$ denotes the activity coefficient at infinite dilution of species 1 in the given solvent. This quantity, or more conveniently $\ln \gamma^{\infty}_{\alpha,1}$ is apparently large when the solubility is small and vice versa. In the formulation of a CAMD problem, one can define the constraint for solubility of substrate 1 as (Abildskov *et al.*, 2010b)

$$\ln \gamma^{\infty}_{\alpha,1} < \epsilon_1 \tag{2.5}$$

for an appropriately chosen ϵ_1. This eliminates the need for solving the non-linear system of equations in Equation (2.3) and is likely to be sufficiently accurate for

screening purposes. However, $\ln \gamma_{\alpha,1}^{\infty}$ still needs to be obtained from a model. Group contribution methods, such as the Universal Functional Activity Coefficient (UNIFAC) method (Hansen *et al.*, 1991) are especially compatible with CAMD algorithms.

2.1.3 Chemical Equilibrium Composition

For a given chemical reaction, a change of reaction medium changes the composition of reactants and products at chemical equilibrium. Under mild reaction conditions, which one usually has for biocatalytic reactions, the equilibrium composition sets a limit to what extent the reactants can be converted into products. This is perhaps the most appreciable feature of non-aqueous biocatalysis, since it for instance allows the process developer to reverse the biological function of hydrolytic enzymes, such as esterases, lipases and proteases (Zaks and Klibanov, 1985). In aqueous solution, these enzymes catalyze the breakdown of esters, lipids and peptides through hydrolysis. In organic media, the reaction equilibria are shifted towards the reactants to such an extent that the opposite reaction is favored, which frequently is the one sought for synthetic applications (Carrea and Riva, 2000). The shift in chemical equilibrium composition can however vary dramatically with a particular choice of organic solvent. For instance, Valivety *et al.* (1991) reported that esterification equilibrium substrate conversions varying over three orders of magnitude, depending on the solvent. This is apparently a key phenomenon to consider in solvent selection. Well-established property models can in many cases be employed to predict the equilibrium composition with satisfactory results.

A chemical reaction can generally be written as

$$|v_1|R_1 + \cdots + |v_{M_r}|R_{M_r} \rightleftharpoons |v_{M_r+1}|R_{M_r+1} + \cdots + |v_{M_r+M_p}|R_{M_r+M_p} \tag{2.6}$$

where R_1, …, R_{M_r} are reactants, R_{M_r+1}, …, $R_{M_r+M_p}$ are products, and v_1, …, $v_{M_r+M_p}$ are stoichiometric coefficients, which by definition are negative for reactants and positive for products. In order to quantify how far a reaction is progressed, one introduces the reaction coordinate ϵ, defined by the differential relation

$$d\epsilon = \frac{dn_1}{v_1} = \frac{dn_2}{v_2} = \cdots = \frac{dn_{M_r+M_p}}{v_{M_r+M_p}} \tag{2.7}$$

where n_i denotes the number of moles of species R_i. The corresponding mole fraction x_i is related to ϵ via

$$x_i = \frac{n_i^{(0)} + v_i\epsilon}{\sum_j \left(n_j^{(0)} + v_j\epsilon\right)} \tag{2.8}$$

where $n_i^{(0)}$ denotes the number of moles present at the reaction start. By minimizing the total Gibbs energy with respect to ϵ, one can derive the equation (Smith *et al.*, 2005)

$$\sum_i v_i \ln x_i\gamma_i = \ln K \equiv \frac{1}{RT}\sum_i v_i G_{\mathrm{f},i}^{\circ} \tag{2.9}$$

which must be satisfied at chemical equilibrium. Here, R, γ_i and $G_{\mathrm{f},i}^{\circ}$ denote respectively the gas constant, the activity coefficient of species R_i and the molar Gibbs

energy of formation of pure species R_i in its standard state at temperature T. The derivation of Equation (2.9) assumes that the reaction takes place in a liquid phase where pressure effects can be neglected. The equilibrium constant K depends on temperature, but not on any other reaction conditions, such as the reaction medium, and can be calculated if $G^{\circ}_{\mathrm{f},i}$ is known for all involved species. The solvent effects are contained in the activity coefficients, which are functions of the reaction mixture composition, and must be obtained from a predictive method, such as UNIFAC (Hansen *et al.*, 1991). Combining Equation (2.8) and Equation (2.9) yields a non-linear equation in ϵ, from which equilibrium composition and product yield can be determined. The approach is also valid when the involved species participate in more than one reaction. In such a situation, one defines a reaction coordinate for each reaction and obtains a system of equations, where each equation is analogous to Equation (2.9) and corresponds to one of the reactions taking place.

A simplification is to treat the reactants and products as dilute by approximating their activity coefficients in Equation (2.9) by the activity coefficients at infinite dilution in the solvent S, $\gamma^{\infty}_{\mathrm{S},i}$, arriving at

$$\sum_i \nu_i \ln x_i = -\sum_i \nu_i \ln \gamma^{\infty}_{\mathrm{S},i} + \ln K \tag{2.10}$$

This treatment is often well-motivated for non-aqueous biocatalytic processes, since mole fractions of reactants and products often are less than 1 %. A solvent-dependent equilibrium constant $\ln K_{\mathrm{S}}$ can be defined by the right-hand side of Equation (2.10)

$$\ln K_{\mathrm{S}} \equiv \ln K - \sum_i \nu_i \ln \gamma^{\infty}_{\mathrm{S},i} \tag{2.11}$$

The larger the value of K_{S}, the higher conversion is obtained at chemical equilibrium. The ratio of such equilibrium constants in two different solvents, S and A can be expressed as

$$\frac{K_{\mathrm{S}}}{K_{\mathrm{A}}} = \prod_i P^{\nu_i}_{\mathrm{SA},R_i} \tag{2.12}$$

where P_{SA,R_i} denotes the S-A partition coefficient for species R_i at infinite dilution, defined as $\gamma^{\infty}_{\mathrm{A},i}/\gamma^{\infty}_{\mathrm{S},i}$, i.e. the ratio between the activity coefficients at infinite dilution for species R_i in solvents A and S.

Halling (1990b) explored predictions of the equilibrium shift based on Equation (2.12). Predictions for esterification in 15 common solvents were reported. The partition coefficients were either taken from experimental data or from the group contribution approach by Rekker and de Kort (1979), which agreed well with the data derived from experiments. Predictions based on UNIFAC (Hansen *et al.*, 1991) were also explored, but the agreement with experimental data was merely qualitative. It is noteworthy that water, which is a by-product of esterification, was treated on a different basis than the other species. Instead of assuming this component to be dilute, it was assumed to be present at a fixed thermodynamic activity, $a_{\mathrm{w}} \equiv x_{\mathrm{w}}\gamma_{\mathrm{w}}$, which was the same in all reaction media. A consequence of this is that the partition coefficient of water does not enter in Equation (2.12) for the equilibrium constant ratio. The rationale behind this assumption was that controlling the water activity

allows to optimize the biocatalyst activity, which was explored in a subsequent study (Valivety *et al.*, 1992b).

Halling (1990b) further demonstrated that the predicted solvent-induced shift of equilibrium position towards the products could be successfully correlated with the solubility of water in the solvent. Attempts to correlate the equilibrium shift with other measures of solvent polarity, such as the octanol-water partition coefficient $\log P$, or the dielectric constant were less successful. These findings were experimentally validated by Valivety *et al.* (1991), who measured equilibrium composition for esterification of dodecanol and dodecanoic acid in different solvents with immobilized Porcine pancreatic lipase as catalyst.

Stamatis *et al.* (2000) attempted similar predictions for transesterification of hexanol and ethyl acetate. They however employed the more rigorous approach determining the equilibrium composition by solution of Equation (2.9). Gibbs energies of formation were taken from literature and the UNIFAC model was used to obtain the activity coefficients as functions of the composition. Experiments were also performed for eleven pure and nine solvent mixtures using immobilized *Candida antarctica* lipase B (CALB) as catalyst, and the measured equilibrium compositions were compared with the predicted ones. The predicted values were in good agreement with the experimental, but the predictions failed to reproduce small variations with solvent present in the experimental results. This was probably due to that in transesterification, an ester reacts with an alcohol to form another ester and alcohol. The products contain therefore precisely the same chemical groups as the reactants, and a group contribution approach thus yields predictions that are nearly independent of solvent.

Predictions of the shift of chemical equilibrium composition can be incorporated in a CAMD screening protocol. It is for this purpose appropriate to consider the solvent-dependent equilibrium constant $\ln K_{\mathrm{S}}$ of Equation (2.11), since its magnitude indicates the equilibrium conversion. For the purpose of ranking solvents with respect to conversion, one can further disregard the term $\ln K$, which is the same for all solvents, and formulate the design constraint as (Abildskov *et al.*, 2010b)

$$\sum_i v_i \ln \gamma_{\mathrm{S},i}^{\infty} < \epsilon_R \tag{2.13}$$

This approach avoids both the need for solving a non-linear equation (Equation (2.9)) and the use of possibly inaccurate $G_{\mathrm{f},i}^{\circ}$ values.

From a prediction of the chemical equilibrium composition, one can however not in general conclude how fast equilibrium is attained. A solvent that gives a high product yield at equilibrium could very well inhibit the biocatalyst and thus slow down or even prevent the reaction from taking place. A favorable shift of equilibrium position does therefore not necessarily mean that a high equilibrium conversion is practically attainable.

2.1.4 Enzyme Activity

In contrast to thermodynamic properties, which only depend on a limited set of state variables, the kinetics of a reaction is governed by a wide range of conditions. For this reason, accurate prediction of kinetic properties is expected to be much more

difficult than prediction of thermodynamic properties. In particular, the catalytic activity could be affected by interactions between the enzyme and the solvent. Due to the complex structure and dynamics of protein molecules, such effects might be difficult to rationalize. The literature nevertheless contains attempts to correlate enzyme kinetics in organic solvents with thermodynamic properties of the solvent. Two examples are described below.

An early study by Laane *et al.* (1987) demonstrated that the catalytic activity of transesterifications catalyzed by fungal lipases from *Candida cylindracea* and *Mucor* was correlated with solvent $\log P$, with P being the octanol-water partition coefficient. The activity showed a sigmoidal dependence on $\log P$, with a transition from low to high activity at around $\log P = 3$. This correlation was however not found for similar measurements with porcine pancreatic lipase as catalyst. The transesterification activity of this enzyme did however correlate with $\log P$ in the study of Valivety *et al.* (1991). Laane *et al.* (1987) suggested that the correlation could be due to that polar solvents (low $\log P$) disrupt the water layer around the enzyme, which is believed to be essential for activity. Non-polar solvents (high $\log P$) on the contrary leave the hydration layer intact. This is generally in line with the observation that enzymes lose their activity upon complete dehydration (Zaks and Klibanov, 1988a). The correlation is however not completely general since other studies have reported solvent-dependent activities that do not obey the correlation (Degn and Zimmermann, 2001).

In a recent study, Abildskov *et al.* (2010b) investigated if the catalytic activity was correlated with the shift of chemical equilibrium composition. Although a general relation between kinetic and thermodynamic properties cannot be expected, it is a plausible hypothesis that the absence of equilibrium is a driving force behind the reaction. The attempt to correlate the initial reaction rates with the value of the reaction coordinate ϵ (see Section 2.1.3) at equilibrium is furthermore straightforward.

In the study, initial rates of transesterification and esterification catalyzed by CALB were analyzed. Measurements of transesterification[1] were reported in the same study (Abildskov *et al.*, 2010b), while data for esterification was taken from measurements performed by Nordblad and Adlercreutz (2008). The equilibrium reaction coordinate ϵ was computed by solution of Equation (2.9), employing UNIFAC (Hansen *et al.*, 1991) to obtain activity coefficients. The results are shown in Figures 2.1 and 2.2(a)-(b).

Figure 2.1 shows that for transesterification, the initial rates in different solvents correlate quite well with predicted equilibrium reaction coordinates. The correlation appears linear, although no known underlying reason suggests this. It seems that ϵ actually can be useful as predictor for the enzymatic activity. There are however a few solvents that divert from the trend, namely dimethyl formamide (DMF), isooctane and methyl t-butyl ether (MTBE). The activity is unexpectedly low in the two former, while it is unexpectedly high in the latter.

For esterification, the equilibrium reaction coordinate was calculated by two alternative routes, namely with or without "water removal". Without "water removal", water is in the solution of Equation (2.9) treated on the same basis as the other

[1]Measurements of transesterification kinetics were performed by M.B. van Leeuwen, C.G. Borieu and L.A.M. van den Broek, Wageningen UR Food & Biobased Research, The Netherlands.

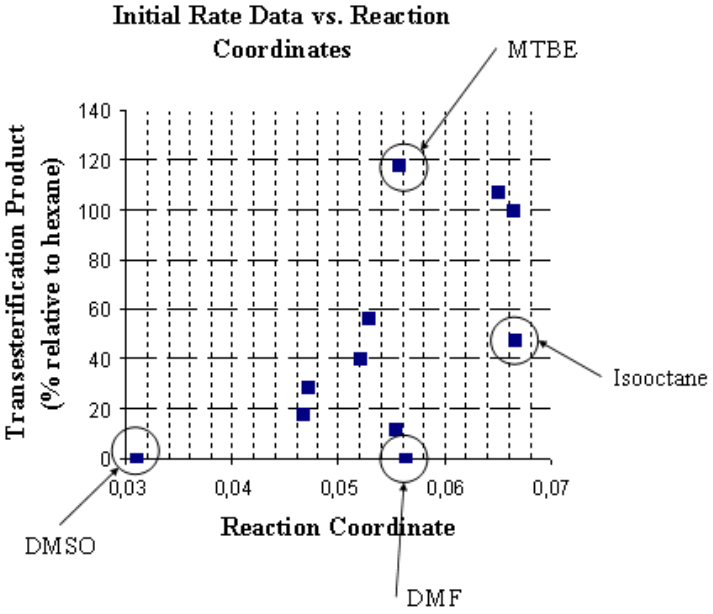


Figure 2.1: Measured initial rates for transesterification of octanol with vinyl laurate in different organic solvents vs. predicted reaction coordinates ϵ at equilibrium. Reproduced from Abildskov *et al.* (2010b).

products and reactants. With "water removal", it is assumed that the reaction proceeds with a fixed thermodynamic water activity. This corresponds to a situation where the water produced during the reaction is immediately removed from the bulk medium. Possibly, a hydration layer on the enzyme surface could act as a buffer adsorbing the produced water. As seen in Figure 2.2(a) there is no apparent correlation between the measured initial rates and ϵ as predicted without "water removal". If, on the other hand, "water removal" is employed, the initial rates appear again to increase with increasing ϵ, as seen in Figure 2.2(b). There are nevertheless a few solvents that the predicted ϵ fails to rank. A higher initial rate was measured in t-butanol than in acetone and 2-butanone, while the predicted ϵ is higher in the two latter solvents. Similar to the case of transesterification, the predicted equilibrium reaction coordinates reproduce the overall trend, but apparently, other factors play a role as well.

The examples discussed above illustrate how solvent effects on reaction kinetics sometimes are correlated with thermodynamic properties. It is however not a surprising fact that medium effects on kinetics cannot be entirely explained in terms of thermodynamics. As will be argued in Section 2.2, molecular modeling can be a useful tool for understanding the effects of physical interactions between the enzyme and the medium. Such phenomena might be relevant for explaining unexpected effects on kinetics, such as the outliers in Figures 2.1 and 2.2(b). Molecular modeling might also be used for instance to validate the assumption behind the "water removal" treatment in the prediction of ϵ.

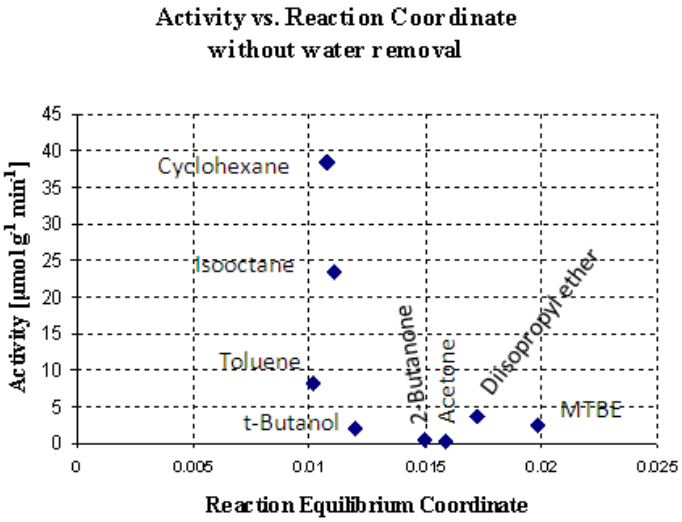
2.1.5 Enzyme Specificity

The approach by Wescott and Klibanov (Wescott and Klibanov, 1993a,b; Ke *et al.*, 1996) attempts to rationalize the solvent dependence of enzyme substrate specificity completely in terms of substrate solvation thermodynamics. It aims to predict the change in the ratio of the apparent specificity constants $k_{\mathrm{cat}}/K_{\mathrm{M}}$ for two substrates, R_1 and R_2, that will accompany a change of solvent. The approach is based on the thermodynamic cycle below, where solvent A is taken as reference medium

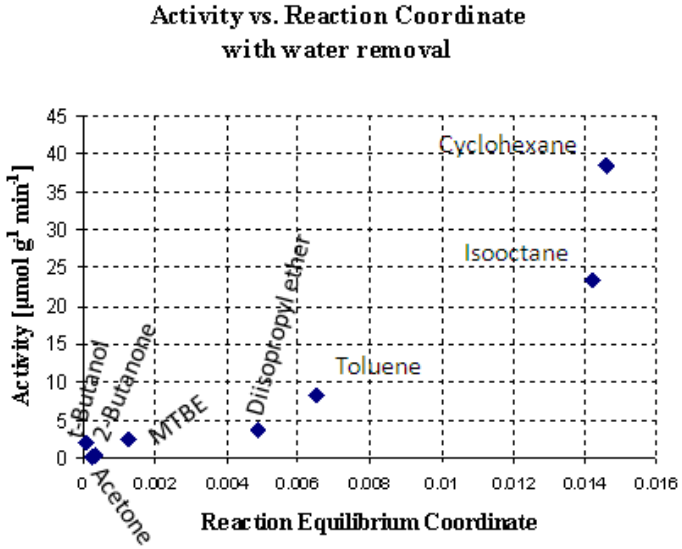
$$\begin{array}{ccc} (E+R_1)_{\mathrm{A}} & \overset{\Delta G^{\ddagger}_{\mathrm{A},R_1}}{\longrightarrow} & (ER_1^{\ddagger})_{\mathrm{A}} \\ {\scriptstyle \Delta G_{\mathrm{SA},E+R_1}} \uparrow & & \downarrow {\scriptstyle \Delta G_{\mathrm{AS},ER_1}} \\ (E+R_1)_{\mathrm{S}} & \overset{\Delta G^{\ddagger}_{\mathrm{S},R_1}}{\longrightarrow} & (ER_1^{\ddagger})_{\mathrm{S}} \end{array} \tag{2.14}$$

where $E + R_1$ denotes the free enzyme and substrate in solution and $ER_1^{\ddagger}$ the transition state, i.e. the least stable atomic configuration along the reaction pathway. This state is reached when the enzyme-substrate complex is formed and covalent modifications of the substrate have started to take place. $\Delta G^{\ddagger}_{\mathrm{S},R_1}$ and $\Delta G^{\ddagger}_{\mathrm{A},R_1}$ denote the molar Gibbs energy costs for forming the transition states in the solvents S and A, respectively. Employing simple transition state theory (TST), one can approximate these free energy costs from the corresponding reaction rates (Laidler, 1987)

$$\Delta G^{\ddagger}_{\mathrm{S},R_1} = -RT\ln\left[\frac{h}{k_{\mathrm{B}}T}\left(\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}\right)_{\mathrm{S},R_1}\right] \tag{2.15}$$



(a)



(b)

Figure 2.2: Measured initial rates for esterification of octanol and acrylic acid in different organic solvents vs. predicted reaction coordinates ϵ at equilibrium (a) without "water removal" or (b) with "water removal", as explained in the text. Reproduced from Abildskov *et al.* (2010b).

where R, h, k_B and T denote the gas constant, Planck constant, Boltzmann constant and temperature, respectively, and $(k_{cat}/K_M)_{S,R_1}$ is the specificity constant for R_1 in the solvent S. An equivalent relation holds between $\Delta G^{\ddagger}_{A,R_1}$ and the specificity constant for R_1 in solvent A, $(k_{cat}/K_M)_{A,R_1}$. $\Delta G_{S,E+R_1}$ is the molar Gibbs energy cost for transferring free enzyme and substrate from solvent S to solvent A, and is expressed in terms of the corresponding S-A partition coefficients, $P_{SA,E}$ and P_{SA,R_1} (see definition in Section 2.1.3), according to

$$\Delta G_{S,E+R_1} = RT\ln P_{SA,E} + RT\ln P_{SA,R_1} \tag{2.16}$$

By combining Equations (2.14)-(2.16) with their counterparts for the competing substrate R_2, Wescott and Klibanov (1993b) derived the relation

$$\ln\left[\frac{(k_{cat}/K_M)_{S,R_2}}{(k_{cat}/K_M)_{S,R_1}}\right] = \ln\frac{P_{SA,R_1}}{P_{SA,R_2}} + \ln\left[\frac{(k_{cat}/K_M)_{A,R_2}}{(k_{cat}/K_M)_{A,R_1}}\right] + \frac{\Delta G_{AS,ER_1} - \Delta G_{AS,ER_2}}{RT} \tag{2.17}$$

They furthermore argued that the last term on the right-hand side can be neglected if both substrates are completely shielded from the solvent, when bound to the enzyme, and arrived at

$$\frac{(k_{cat}/K_M)_{S,R_2}}{(k_{cat}/K_M)_{S,R_1}} = \text{constant} \times \frac{P_{SA,R_1}}{P_{SA,R_2}} = \text{constant} \times \frac{\gamma^{\infty}_{S,R_2}}{\gamma^{\infty}_{S,R_1}} \tag{2.18}$$

where the constant is independent of the solvent S. In the last step, the relation between the partition coefficients and the activity coefficients at infinite dilution of the substrates in the current solvent, γ^{∞}_{S,R_1} and γ^{∞}_{S,R_2}, have been employed. The partition coefficients or values of γ^{∞} can be obtained from literature or from well-established predictive methods, such as UNIFAC (Hansen *et al.*, 1991). Equation (2.18) provides a simple and efficient route to predict the change in specificity upon changing the solvent. For the transesterification reaction of either N-Ac-L-Ser-OEt or N-Ac-L-Phe-OEt with 1-propanol catalyzed by Subtilisin Carlsberg, the ratio of the specificity constants for the two ester substrates was shown to correlate well with Equation (2.18) for a wide range of solvents, using measured partition coefficients (Wescott and Klibanov, 1993b).

A limitation of the Wescott-Klibanov model is that it yields predictions that are independent of the enzyme. It does not predict any solvent effect on enantioselectivity, since the partition coefficients or the infinite-dilution activity coefficients are similar for two enantiomers. If furthermore fails to predict solvent effects on prochiral, regio- and chemoselectivity, since in those cases, the competition is between different chemical modifications of the same substrate and not between different substrates. The assumption that the substrate is completely shielded from the solvent is relaxed in a refined version of the model (Ke *et al.*, 1996). It is assumed that in the transition state, part of the substrate is "sticking out" of the pocket and is accessible to the solvent. Equation (2.18) is replaced by

$$\frac{(k_{cat}/K_M)_{S,R_2}}{(k_{cat}/K_M)_{S,R_1}} = \text{constant} \times \frac{\gamma'^{\infty}_{S,R_2}}{\gamma'^{\infty}_{S,R_1}} \tag{2.19}$$

Here, γ'^{∞}_{S,R_1} and γ'^{∞}_{S,R_2} are defined as the activity coefficients of the parts of substrates R_1 and R_2 that are inaccessible to the solvent when the corresponding substrate is bound to the enzyme. They are not actual physical properties and can

therefore not be measured experimentally, but can be defined and computed using group contribution methods. In order to determine which groups of the substrate are shielded from the solvent, molecular modeling is employed. The transition state is approximated by the tetrahedral intermediate which is modeled using the crystal structure coordinates and building the substrate covalently bound to the active site. The structure is relaxed through a series of energy minimizations and short molecular dynamics (MD) simulations. The solvent accessible surface area (SASA) of the substrate groups is calculated and used to determine which groups are inaccessible to solvent in the relaxed structure. This approach successfully predicted the solvent dependence of pro-chiral selectivity for esterifications catalyzed by crystalline γ-Chymotrypsin and Subtilisin Carlsberg, but failed to predict the corresponding selectivity of lyophilized and acetone-precipitated Chymotrypsin (Ke *et al.*, 1996). The authors suggested that the failure was due to the fact that the enzyme structure was non-native in those particular preparations.

In a later study, Colombo *et al.* (1998) evaluated the approach outlined above for rationalizing solvent effects on subtilisin enantioselectivity. They considered transesterifications of *sec*-phenetyl alcohol and *trans*-soberol with vinyl acetate or vinyl butyrate as model reactions, and employed either lyophilized or crystalline subtilisin as catalyst. They found the predictions of Equation (2.19) to be in poor agreement with their measurements.

The method of Wescott and Klibanov is due to its simplicity useful in medium engineering for reactions where enzyme specificity is of importance (Carrea and Riva, 2000). The original version of the model (Equation (2.18)) is easily included in the initial screening of solvents using a CAMD algorithm. A limitation of the approach is however that it predicts enzyme specificity, but gives no information about the solvent effects on the absolute activities. A solvent that greatly favors substrate R_1 over R_2 could very well be slowing down the absolute reaction rate substantially and thus be a poor solvent. The results of Colombo *et al.* (1998) furthermore shows that substrate solvation is not the only way in which the solvent affects enzyme specificity. Other phenomena that might be of importance are solvent-induced conformational changes of the enzyme and binding of solvent molecules to the active site.

2.1.6 Enzyme Stability

The stability of an enzyme refers to its ability to remain active over long time. The more specific term thermo-stability refers to how well enzyme activity is retained when exposed to high temperature. As emphasized by Klibanov (2001), there are two different phenomena that might cause enzyme inactivation. The first is unfolding, which typically is instantaneous and reversible. The second involves covalent modifications of the protein molecules, due to interactions with the solvent. This type of inactivation occurs over longer time and is irreversible. Both phenomena are promoted by e.g. high temperature. Most investigations indicate that enzymes are more thermo-stable in organic solvents than in water (Zaks and Klibanov, 1984; Griebenow and Klibanov, 1996). Partly, this has been attributed to that conformational flexibility is lower in organic solvents than in water, and that the free energy barriers that prevent unfolding therefore are more difficult to overcome in organic media. It has also been argued that the higher stability observed in some organic

solvents is due to that these particular solvents are less reactive than water, and are less likely to induce covalent modifications of the enzyme.

Interestingly, Griebenow and Klibanov (1996) observed by means of FTIR spectroscopy that lysozyme and subtilisin maintained their native structure in nearly pure acetonitrile, tetrahydrofuran and 1-propanol. When a substantial amount of water was added to either of these solvent, the suspended enzymes were nevertheless seen to unfold. The authors proposed that the native structures of the two enzymes are thermodynamically unstable in organic media, but kinetically stable, due to the low conformational mobility. In the presence of some water, the enzyme flexibility increases, and the kinetic barrier that prevents unfolding can be overcome.

There seems to be no study which attempts to quantitatively correlate enzyme stability with solvent properties.

2.1.7 Other Solvent Properties

The simplest properties to be considered in solvent selection for biocatalysis are those that only depend on the solvent, and neither on substrates, products or the enzyme preparation. Any list of properties of this type that are relevant for solvent selection is necessarily non-exhaustive. It is however worthwhile to note that predictive models are available for a range of such properties, some described by Poling *et al.* (2007). For safe handling of the solvent, it is e.g. relevant to consider the toxicity and flammability.

The solvent toxicity can be quantified e.g. by the LC_{50} measure, which is defined as the concentration of solvent present in aqueous solution that during a specific time interval causes mortality in 50% of a fish population of a particular species. A group contribution method for predicting this measure has been developed by Martin and Young (2001).

Flammability depends on the solvent flash point, which is the temperature at which the density of the vapor is high enough to form an ignitable mixture with air. The flash point is typically correlated with the solvent boiling point (Butler *et al.*, 1956), which e.g. can be predicted by the group contribution method by Constantinou and Gani (1994).

2.1.8 Summary

Modeling enzyme-independent parameters is computationally efficient and relatively straightforward. The limitation of CAMD and enzyme-independent models is not surprisingly encountered for parameters that are linked directly to the enzyme, such as activity, specificity and stability. Attempts to correlate such properties with solvent properties have only been partially successful. In Sections 2.1.4 and 2.1.5, it was proposed that molecular modeling could help rationalizing anomalistic results. Molecular modeling methods are however very computationally demanding, and can therefore not be used directly in a CAMD screening. If employed in a solvent selection strategy, they should constitute a second step and only be used to evaluate solvent candidates indentified by the simple, enzyme-independent screening.

2.2 Molecular Modeling of Enzymes in Non-aqueous Media

Molecular modeling generally refers to a collection of methods used to study the behavior of systems of molecules. One of the principal strengths of these methods is that they provide structural and dynamical information with a level of detail that is difficult, if not impossible, to obtain by experimental techniques. Molecular modeling is well-suited for interpreting molecular-scale phenomena in terms of fundamental physical interactions, and therefore a useful tool for understanding the behavior of proteins in non-aqueous media. This section gives a brief description of two molecular modeling methods, namely classical molecular simulation and quantum mechanical methods. Focus is placed on how these methods can contribute to the understanding of enzyme function, and an overview of previous molecular modeling studies of enzymes in non-aqueous media is given. Before shifting the focus to these methods, it is appropriate to recall some experimental observations of the behavior of proteins in non-aqueous media.

It is well established that enzymatic activity and selectivity can be promoted or inhibited by interaction with water and organic solvent molecules. How this happens is however not fully understood. Commonly, solvent effects on activity and selectivity are attributed to thermodynamic factors, such as those described previously in this chapter. Other hypotheses that have been discussed in the literature include

1. Solvent impacts the protein structure and conformational flexibility which in turn affects catalytic activity.

2. Water or organic solvent molecules bind to the active site acting as inhibitors.

3. Solvents differ in their ability to stabilize the highly polarized transition state which is temporarily formed during the covalent steps of catalysis.

Several studies have demonstrated that enzymes typically retain their tertiary structure upon transfer from water to organic solvent (Fitzpatrick *et al.*, 1993, 1994; Yennawar *et al.*, 1994; Griebenow and Klibanov, 1996). They do however loose conformational flexibility. There are many indications that flexibility is a key factor in protein function Fersht (1999). The precise role of flexibility is however not understood in detail and is likely to be different for different proteins. A reasonable hypothesis is that flexibility around the active site allows the enzyme to adapt to the shape of a substrate and thus bind it efficiently. Several studies have attempted to establish a direct correlation of enzyme flexibility and catalytic activity. Such a correlation was observed by Affleck *et al.* (1992) (electron spin resonance of active site spin-labeled subtilisin), Broos *et al.* (1995) (time-resolved fluorescence anisotropy of α-chymotrypsin and subtilisin), and Watanabe *et al.* (2004) (electron spin resonance of active site spin-labeled subtilisin). In those studies, the enzymes were suspended in dry organic solvents. By adding water, the flexibility was gradually increased, and so was the catalytic rate.

Several researchers have argued that water and organic solvent molecules (in particular alcohols) can bind to the active site, making it inaccessible to the substrate (Valivety *et al.*, 1993; Bovara *et al.*, 1993; Martinelle and Hult, 1995; Graber *et al.*,

2003, 2007; Foresti *et al.*, 2009). The claim has been supported by measurements of the apparent Michaelis-Menten constants, k_{cat} and K_{m}. K_{m} is often seen to decrease with increasing water activity, suggesting that water molecules block the active site. Similar inhibitory effects due to organic solvent molecules have been shown by kinetic measurements in a solid/gas reactor where the amount of organic solvent present can be fine-tuned (Graber *et al.*, 2007).

During catalysis, a temporary chemical bond is often formed between the substrate and the active site of the enzyme. The process of binding involves the formation of a highly polarized transition state (Fersht, 1999). If the polarized region is accessible to solvent, water and organic solvent molecules might stabilize the transition state and thus promote the turnover rate. In particular, water and polar organic solvents might be expected to better stabilize the transition state than non-polar solvents. Xu *et al.* (1994) and Kim *et al.* (2000) proposed this effect to be of significance for the activity of enzymes in organic solvents.

Molecular modeling methods are useful for examining hypotheses such as those mentioned above. The remainder of this section is therefore devoted to describe a selected set of such approaches and their contributions.

2.2.1 Classical Molecular Simulation

Force Fields In classical molecular simulation, an empirical model termed *force field* is used to describe interactions between atoms. In the case of *all atom* simulation, each atom in the studied system is represented by its position in three-dimensional space. The force field is virtually an expression for the total configurational energy of the system in terms of the atom coordinates. It typically has the form (Ponder and Case, 2003)

$$U_{\mathrm{config}} = U_{\mathrm{bond}} + U_{\mathrm{angle}} + U_{\mathrm{dihedral}} + U_{\mathrm{improper}} + U_{\mathrm{LJ}} + U_{\mathrm{Coulomb}} \tag{2.20}$$

where the terms on the right-hand side respectively denote the configurational energy due to chemical bond stretching, bond angle bending, dihedral angle twisting, improper angle displacement (deviation from planarity of sp^2 bonds), Lennard-Jones and electrostatic interactions. The first four terms are referred to as intramolecular terms as they account for interactions between atoms belonging to the same molecule and ensure that proper molecular geometry is maintained. The last two terms are referred to as intermolecular terms as they account for interactions between atoms of different molecules, in addition to interactions between atoms belonging to the same molecule, but separated by at least four bonds.

Several force fields have been developed for simulation of biomolecules. The ones most widely used include AMBER (Assisted Model Building and Energy Refinement) (Cornell *et al.*, 1995), CHARMM (Chemistry at Harvard Molecular Mechanics) (MacKerell Jr. *et al.*, 1998), GROMACS (Groningen Machine for Chemical Simulations) (van der Spoel *et al.*, 2005) and OPLS (Optimized Potential for Liquid Simulations) (Kaminski *et al.*, 2001). The potential energy function of the CHARMM force field, which is employed within this PhD work, is described in more detail in Appendix A.

It is noteworthy that classical molecular simulation is not restricted to all atom simulation. Coarse-grained force fields, in which the fundamental units are groups

of atoms rather than individual atoms, are available for a wide range of purposes.

Monte Carlo and Molecular Dynamics Classical molecular simulation generally uses either the Monte Carlo (MC) or the molecular dynamics (MD) method (Allen and Tildesley, 1987) to simulate a molecular system. With MC, the system is propagated by means of random moves. The moves can e.g. comprise translation and rotation of molecules, torsional rotation of dihedral angles and re-growth of a branch of a molecule. Prior to the simulation, one selects which moves to be utilized and assigns them probabilistic weights. At each simulation step, one of these moves is randomly selected and used to update the system. The change of configurational energy, ΔU_{config} is calculated, and the executed move is accepted with the probability

$$P_{\text{acc}} = \begin{cases} \exp\left(-\frac{\Delta U_{\text{config}}}{k_{\text{B}}T}\right), & \Delta U_{\text{config}} > 0 \\ 1, & \Delta U_{\text{total}} < 0 \end{cases} \tag{2.21}$$

If not accepted, the move is rejected and the system is reverted to its previous configuration. This procedure ensures that after an initial equilibration period, the simulation samples the "canonical" NVT ensemble of the system (constant molecule number N, constant volume V and constant temperature T). By introducing MC moves that allow the simulation box to grow or shrink, one can instead sample the "isothermal-isobaric" NPT ensemble (constant molecule number N, constant pressure P and constant temperature T). With MC moves that insert or remove molecules in the box, one can sample the "grand-canonical" μVT ensemble (constant chemical potential μ, constant volume V and constant temperature T).

With MD, the system is propagated deterministically by numerical integration of Newton's equations of motion. These are most elegantly expressed in Hamiltonian form (Goldstein, 1980)

$$\begin{aligned} \frac{d\boldsymbol{r}_i}{dt} &= \frac{\boldsymbol{p}_i}{m_i}, i = 1, \ldots, M \\ \frac{d\boldsymbol{p}_i}{dt} &= -\nabla_{\boldsymbol{r}_i} U_{\text{total}}(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M), i = 1, \ldots, M \end{aligned} \tag{2.22}$$

where t denotes time, $\boldsymbol{r}_i$, $\boldsymbol{p}_i$ and m_i denote respectively position, momentum and mass of atom i, $\nabla_{\boldsymbol{r}_i}$ denotes the gradient with respect to the position of atom i, and M denotes the number of atoms in the system. By numerical integration of Equation (2.22), one obtains a system trajectory from which configurations can be sampled for analysis. According to the ergodic hypothesis, these configurations are after an initial equilibration distributed according the micro-canonical NVE ensemble (constant molecule number N, constant volume V and constant total energy E). The NVT ensemble can be sampled by employing a thermostat, which simulates the effects of an external heat bath at temperature T. Likewise, the NPT ensemble can be sampled using also a barostat simulating the effect of a piston applying a pressure P to the system. A detailed description is beyond the scope of this thesis, and the reader is referred to the books of Allen and Tildesley (1987) or Frenkel and Smit (2002).

In both MC and MD, structural and thermodynamical properties of the system are obtained via functions of the sampled configurations averaged over the simulation

trajectory. Simulation length is usually chosen such that the properties are obtained with acceptable statistical precision. Simulation length depends on the system and the property of interest.

For simulation studies of proteins in explicit solvent, it is far more common to use MD than MC for two reasons. For an MC simulation to be efficient, it is necessary to carefully select update moves that are appropriate for the studied system, and to assign them appropriate probabilistic weights. An MC algorithm is therefore often implemented with a particular application in mind. MD programs can in contrast be applied to systems of very different character without any need for adapting the algorithm. For this reason, MD software packages are in more widespread use than their MC counterparts (Maginn, 2009). Another reason for preferring MD over MC is that the former method is easier to implement for execution on parallel computing machines. Simulations of proteins in explicit solvent comprise large systems, typically containing at least 25000 atoms, and the simulations are with current technology very time-consuming when executed on a single CPU.

2.2.2 Application to Enzymes in Non-Aqueous Media

Table 2.1 summarizes a number of important MD studies of proteins in non-aqueous media. In none of these studies was the protein seen to undergo large conformational changes induced by the solvent. The studied protein was however seen to be less flexible in organic media than in water in nearly all of the studies.

Important exceptions to this are MD studies of lipases that possess known lid regions. In those studies, it was typically observed that the lid region underwent a conformational change in media of low polarity. The flexibility was therefore higher in those media (Norin *et al.*, 1994; Peters *et al.*, 1996a,b, 1997; Jääskeläinen *et al.*, 1998; Tejo *et al.*, 2004; Cherukuvada *et al.*, 2005; James *et al.*, 2007; Trodler *et al.*, 2009).

This section focuses on studies of enzymes in non-aqueous media and closely related studies with particular emphasis on studying the medium effects on structure and flexibility. A broader review of proteins and peptides in non-aqueous media is given by Roccatano (2008).

Early Studies Early studies of proteins in organic media were carried out by Hartsough and Merz (1992, 1993) who simulated Bovine pancreatic trypsin inhibitor (BPTI) in pure chloroform and water by MD simulations. They observed the protein to be less flexible in the organic solvent. Greatest differences were seen for the side chains of residues at the protein surface. In water, these side chains were extended out into the solvent while they "fell back" onto the protein surface in chloroform resulting in a lower flexibility. Backbone flexibility was also lower in chloroform, although not as sensitive as that of the side chains. The differences in flexibility were ascribed to that the number of protein-protein hydrogen bonds was higher in chloroform than in water. The number of protein-solvent hydrogen bonds was on the other hand higher in water. The authors did not observe any significant backbone conformational changes during the 150 ps of simulation, which probably is too short to make such observations in any case.

Similar observations were made for *Rhizomucor miehei* lipase (RML) by Norin

Table 2.1: Summary of previous MD studies of proteins in non-aqueous media.

Reference	Enzyme	Solvent
Hartsough and Merz (1992, 1993)	BPTI	Chloroform, water
Toba and Merz (1996)	γ-chymotrypsin	Hexane, water
Toba and Merz (1997)	Subtilisin E	DMF, water, DMF/water mixtures
Norin *et al.* (1994)	RML	Methyl hexanoate, vacuum, water
Zheng and Ornstein (1996a,b,c)	Subtilisin Carlsberg	Acetonitrile, carbon tetrachloride dimethyl sulphoxide (DMSO), water
Peters *et al.* (1996a)	RML	Cyclohexane, methyl hexanoate, water
Soares *et al.* (2003)	Cutinase, Ubiquitin	Hexane, water
Tejo *et al.* (2004)	*Candida rugosa* lipase	Carbon tetrachloride, water
Yang *et al.* (2004)	Subtilisin BPN'	Acetonitrile, octane, tetrahydrofuran, water
Cherukuvada *et al.* (2005)	*Pseudomonas aeruginosa* Lipase	Water and octane/water mixtures
Micaêlo *et al.* (2005)	Cutinase	Hexane, water
James *et al.* (2007)	*Candida rugosa* lipase	Hexane/water, octane/water, decane/water mixtures
Micaêlo and Soares (2007)	Cutinase	Acetonitrile, diisopropyl ether, ethanol hexane, 3-pentanone, water
Micaêlo and Soares (2008)	Cutinase	Ionic liquids
Díaz-Vergara and Piñeiro (2008)	TcTIM	Decane, water, decane/water mixtures
Trodler and Pleiss (2008)	CALB	Chloroform, cyclohexane, isopentane methanol, toluene, water
Cruz *et al.* (2009)	Subtilisin Carlsberg	Acetonitrile, water
Trodler *et al.* (2009)	*Burkholderia cepacia* lipase	Toluene, water
Branco *et al.* (2009)	CALB	Argon/water gas mixture

et al. (1994) who found that the enzyme flexibility was lower in methyl hexanoate than in water. The authors reported the number of protein-protein hydrogen bonds to be larger in methyl hexanoate. They furthermore argued that the lower flexibility also was due to the fact that the electrostatic protein-protein interactions were stronger in methyl hexanoate, which do not screen electrostatic interactions as effectively as water does. An interesting event was the opening of the lid covering the active site, which took place in methyl hexanoate but not in water.

The studies above focused on the differences in protein behavior in an organic solvent and water, which are two extremes in terms of polarity. Zheng and Ornstein (1996a,b,c) took one step further and considered water and several organic solvents, namely acetonitrile, carbon tetrachloride and dimethyl sulphoxide (DMSO). The protein studied was Subtilisin Carlsberg, which in contrast to BPTI is an enzyme that frequently is employed in non-aqueous biocatalysis. The authors furthermore recognized the relevance of including some water molecules in the organic solvent simulations, since enzymes are inactivated upon complete dehydration (Zaks and Klibanov, 1988a). The simulations were thus carried out with the "crystal water" included, i.e. those water molecules that bind to the protein during crystallization and thus are resolved in the structure determination. The flexibility of Subtilisin, which was characterized by the root-mean square fluctuation (RMSF), was reported to be highest in acetonitrile and lowest in carbon tetrachloride, while the water and DMSO simulations yielded no significant difference. These results were unexpected, since the number of protein-protein hydrogen bonds still was smaller in water than in acetonitrile as well as in DMSO. No explanation for the differences in flexibility was however provided. The authors did observe that some of the crystal waters during the course of simulation were stripped from the protein surface and mixed with the bulk solvent. This occurred in the polar solvents acetonitrile and DMSO, but not in the non-polar solvent carbon tetrachloride.

Toba and Merz (1996) simulated γ-chymotrypsin in hexane and investigated the differences arising from two different hydration levels. The number of water molecules included was either 50 or 444, where the latter corresponded to a monolayer of water molecules around the protein. The authors reported that the root-mean square deviation (RMSD) measured from the crystal structure decreased slightly with increasing hydration. They also investigated the impact of distributing the 50 water molecules differently around the protein surface in the initial configuration. This was accomplished by starting from two different crystal structures whose crystal water molecules were at different locations. Some impact was seen on the flexibility, in particular that of the surface loops of the protein. The same authors simulated Subtilisin E and a mutant of the same enzyme, previously designed by protein engineering. Simulations were carried out in water, DMF and a 60/40 mixture of these two solvents (Toba and Merz, 1997). In the mutant, which was experimentally observed to be more catalytically active in DMF/water mixtures than the wild type, several negative amino acids at the surface had been substituted for neutral ones. The authors argued that this would make the protein more compatible with DMF, which cannot stabilize negatively charged residues, which would explain the higher activity. The simulations showed that for the wild type, the RMSF decreased as water was replaced by DMF or DMF/water. The same trend was observed for the mutant. The flexibility of the mutant in DMF/water mixtures was interestingly

similar to that observed for the wild type in water. The authors proposed that the mutant's high activity in DMF/water mixtures could be attributed to that in those media, the dynamics of the mutant was similar to that of the wild type in pure water which was considered as the "native" medium of the wild type.

Influence of Solvent Properties While the studies mentioned above provided elementary insights into protein dynamics in organic solvents, the simulations were rather short and seemingly never longer than one nanosecond. This did thus not allow for larger backbone movements to take place. Little attention was given to quantitative comparisons of protein structure and dynamics obtained from simulations in different organic solvents. Zheng and Ornstein attempted such comparisons in their work, but as stated above, their results were difficult to rationalize. If MD simulations are to be useful in solvent selection for biocatalysis, it is crucial that comparisons of this kind are feasible and able to generate reliable results. A step towards this goal was taken by Trodler and Pleiss (2008) who studied *Candida antarctica* lipase B (CALB) in water and five organic solvents, namely chloroform, cyclohexane, isopentane, methanol and toluene. The 286 crystal water molecules were included and multiple simulations were carried out of each system in order to establish the statistical significance of the results. The flexibility was measured as the sum of the B-factors determined in the simulations, and this measure was shown to be well correlated with solvent dielectric constant and anti-correlated with hydrophobicity, given as $\log P$. These trends were attributed to the dynamics of the crystal water molecules. The term "slowly exchanged water" was introduced for water molecules which remained bound to the same site at the protein surface throughout the simulation. The higher $\log P$ of the solvent, the more slowly exchanged water molecules were observed. In the hydrophilic solvents the crystal water molecules were more prone to move around or be stripped from the surface. The authors thus concluded that CALB flexibility was mainly dependent on the dynamics of the hydration layer and that the solvent affected the flexibility mainly by interacting with the hydration layer.

Such a conclusion raises however a fundamental question, since the size and behavior of the hydration layer depends not only on the solvent, but also on how much water was at the enzyme surface at the start of the simulation. Including the crystal water is a convenient approach, but results in different hydration levels depending on the solvent. In polar solvents, some of the crystal water is stripped from the enzyme surface and mixes with the bulk medium, which thus no longer is a pure organic liquid, but an aqueous/organic mixture. The number of water molecules present in the simulation box is thus an additional parameter that might have an impact on the calculated properties. It is reasonable to ask how sensitive such properties are to the hydration level.

Effect of Hydration Soares *et al.* (2003) carried out MD simulations of two proteins, namely ubiquitin and *Fusarium solani* cutinase in pure water and hexane. The simulations in hexane were carried out with different water contents, ranging from 0 to 25% water weight per protein weight (w/w). Protein properties did indeed correlate with the hydration level. The RMSD for the two proteins was calculated with respect to an average structure obtained from simulations of the two proteins

in pure water. For cutinase the RMSD value showed in hexane a "U"-shaped dependence on the hydration level with a minimum at 10% (w/w). The authors pointed out that the catalytic activity of cutinase was maximal at precisely this hydration level and suggested that this could be explained by the maximal structural resemblance with the protein in an aqueous environment. The total RMSF increased with increasing hydration level for both proteins, and did in fact surpass the RMSF obtained in pure water. Interestingly, the simulations carried out in pure water yielded RMSF values similar to the ones obtained in hexane at 10% (w/w).

A similar study was conducted by Díaz-Vergara and Piñeiro (2008) who simulated *Trypanosoma cruzi* triosephosphate isomerase (TcTIM) for 40 ns in pure water and decane at varying hydration levels. They observed as well that the total RMSF increased with increasing hydration level. From the simulations carried out in pure water, RMSF values were obtained that were similar to the ones obtained in decane at a low hydration level, but higher than the ones obtained in pure decane. Hexane and decane are non-polar organic solvents which mix poorly with water. Consequently, both Soares *et al.* (2003) and Díaz-Vergara and Piñeiro (2008) observed that all the water remained near the protein surface throughout the simulations. Soares *et al.* (2003) reported however that as the number of water molecules in the hydration layer increased, the RMSF values of these water molecules increased meaning that they became more free to move around within the hydration layer.

Yang *et al.* (2004) investigated the effects of hydration in non-polar as well as polar organic solvents. Surfactant-solubilized Subtilisin BPN' was simulated in acetonitrile, octane, and tetrahydrofuran with either precisely the 186 crystal water included, or with 846 water molecules which is just enough to form a monolayer around the protein-surfactant complex. In the simulations of about 5 ns, no significant differences in structure and flexibility could be observed in the different solvents or hydration levels. It was observed that in octane, all water molecules remained near the protein surface. In tetrahydrofuran, 15–20 water molecules were stripped from the protein surface, while in acetonitrile, a significant amount of water mixed with the bulk solvent. The average water density in the active site region was highest in octane, followed by tetrahydrofuran and acetonitrile. The surface water molecules were more mobile and free to move around on the surface in the octane simulations than in the polar solvent simulations, where the surface water molecules mostly were less mobile and located at specific sites. In the acetonitrile simulations, the water molecules around the active site were to some extent replaced by acetonitrile molecules. These molecules were also seen to penetrate the protein structure at other locations. The simulations carried out with 186 water molecules showed the same qualitative trends as those carried out with 846 water molecules, and the authors did not devote much of the discussions to compare the two hydration levels for each solvent.

In the extensive study of Micaêlo and Soares (2007), cutinase was simulated in five organic solvents, namely acetonitrile, diisopropyl ether, ethanol, hexane and 3-pentanone. With each solvent, simulations were carried out at 7–12 different hydration levels, ranging from 5 to 100%, referring to the w/w ratio of the total amount of water and the protein weight. The more polar the solvent was, the fewer water molecules were in the simulations retained at the protein surface, which was consistent with the findings of Yang *et al.* (2004). The authors plotted the average

number of water molecules at the surface versus the total amount of water molecules and pointed out the resemblance with adsorption isotherms. Several aspects of the behavior of the surface water were analyzed. At low hydration, water was located at specific sites on the protein. These sites were found to be independent of the organic solvent. The water molecules at the surface were mainly isolated or organized in small clusters. The number of clusters increased with increasing hydration level until a critical level was reached. The number of clusters was insensitive to further increase of hydration level, which indicated that all binding sites were occupied, and that additional water joined existing clusters instead of forming new ones. In hexane, diisopropyl ether and 3-pentanone this critical level was at a total water content of 25% (w/w). In ethanol and acetonitrile, the number of clusters increased slower than in the non-polar solvents and did not reach saturation. This was probably due to that fewer water molecules were retained on the surface in those solvents. The dynamics of the surface water was characterized by residence times of water molecules at the protein surface. The residence times decreased with increasing water content and with solvent polarity. Cutinase structure and dynamics were analyzed by computing the average RMSD with respect to the crystal structure. For each organic solvent, this quantity was plotted versus the system water content, resulting in rather noisy plots, some nevertheless having distinguished local minima. In hexane, the minimum was observed at a water content of 7.5% (w/w), consistent with the previous study of Soares *et al.* (2003). In the solvents of higher polarity, the minimum was seen at higher total water contents, which was attributed to the observation that in those solvents, the fraction of water actually located around the protein is smaller than in hexane. The authors hypothesized that the conditions yielding minimal RMSD would correspond to similar water activity, although the latter was not assessed. They further suggested that low RMSD might promote the activity, which for cutinase in organic solvents is known to have a bell-shaped dependence on hydration (Vidinha *et al.*, 2003). Interestingly, the same authors reported in another MD study that cutinase in the ionic liquid [BMIM][PF_6] showed a similar RMSD minimum at a water content of 5-10% (w/w) (Micaêlo and Soares, 2008). This resembles the situation in hexane, which seems contradictory, since the authors reported that very few water molecules were retained at the protein surface in simulations with ionic liquids.

Recently, Cruz *et al.* (2009) simulated subtilisin Carlsberg in water, and acetonitrile, with or without the crystal water. Long simulations, i.e. 90 ns, were carried out in acetonitrile. In this solvent, significant structural changes were reported, and the RMSD of C_α atoms with respect to the crystal structure became as high as 6 Å. RMSF values were furthermore higher than in water. An interesting observation was that a structural change occurring in the organic solvent simulations opened up a path to the protein core. This allowed acetonitrile molecules to penetrate deep into the protein, which could be a start for unfolding.

Explicit evaluation of the water activity From the studies summarized above, one can conclude that protein hydration is a crucial parameter in simulation studies of enzymes in non-aqueous media, as several trends observed in these studies have been attributed to the behavior of the hydration water, which is affected by the organic solvent. The hydration level needs to be controlled, or at least carefully

considered if structural and dynamical protein properties are to be reliably obtained from simulation. It is however not obvious how to control the hydration level in simulations. Including a fixed number of water molecules might be inappropriate, since it leads to different hydration levels in different solvents, depending on the solvent polarity. Probably, the most rigorous approach to hydration level control is to fix the water activity of the medium, since in several experimental studies, this parameter has been the key to interpreting results for kinetic parameters (Halling, 1989, 1990b; Valivety *et al.*, 1992b,a; Halling, 1994; Bell *et al.*, 1997).

In the study of CALB conducted by Branco *et al.* (2009), the hydration level was analyzed in terms of the water activity. The medium was however not an organic solvent but a gaseous mixture of water and argon, modeling the carrier gas of a solid/gas reactor. Several simulations were carried out with different amounts of water present, and the bulk water activity was evaluated via the partial pressure of water molecules not associated with the enzyme, assuming the medium to be an ideal gas mixture. At low water activities, the water molecules on the enzymes surface were isolated or organized in small clusters, binding at specific sites on the protein. The number of clusters increased with increasing water activity, until the latter reached a value of 0.5. Beyond that, the number of clusters decreased as the clusters percolated to form a layer. The overall structure and flexibility of CALB seemed uncorrelated with the water activity, while two local segments at the active site entrance were identified, whose flexibility increased with increasing hydration.

There seems to be no studies investigating the effects of different water activities in the presence of organic solvent. For a liquid medium, it is not possible to evaluate the water activity in terms of the partial pressure as done by Branco *et al.* (2009). Seemingly, there are no well-established methods for controlling or assessing the water activity in simulations of such systems. The development of such a method, and the establishment of the water activity dependence of protein properties in the presence of organic solvent is presently a key challenge in this field of study. This might strengthen the impact of computational studies since it might allow for more quantitative comparisons with experimental measurements of enzyme kinetics, especially those carried out at fixed water activities.

2.2.3 Quantum Mechanical Approaches

In a classical molecular simulation utilizing a force field, chemical bonds are predefined and remain fixed during the course of simulation. In order to study formation or breaking of chemical bonds, which is of central importance in catalysis, quantum mechanical (QM) methods are usually employed. For enzymatic reactions, QM methods are commonly applied to determine reaction pathways, investigate the role of individual residues, rationalizing enzyme selectivity or to estimate reaction rates.

For the purpose of studying chemical reactions, the starting point is in most cases the eigenvalue equation (Engel and Hehre, 2005)

$$\hat{H}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)\Psi = E(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)\Psi \tag{2.23}$$

which is derived from the Schrödinger equation for the electronic wavefunction Ψ. The Born-Oppenheimer approximation is employed, meaning that the nuclei are treated as static point charges and that the Hamiltonian $\hat{H}$ is a function of the

nuclear coordinates $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$. The smallest eigenvalue E, i.e. the ground state energy, is as well a function of the nuclear coordinates and defines a potential energy surface, which in the Born-Oppenheimer approximation governs the dynamics of the nuclei. As given, Equation (2.23) can neither be solved analytically or numerically without the use of further approximations. Numerous approaches varying in accuracy and computational complexity are established and fall roughly into three categories. The first category includes the *ab initio* methods, which typically employ a variational method to find an approximate solution to Equation (2.23). This category includes the most accurate (post-Hartree-Fock, e.g. Møller-Plesset perturbation theory), but also most computationally expensive methods. The second category consists of the density functional theory (DFT) methods (e.g. B3LYP), in which the problem of Equation (2.23) is reformulated in terms of the electron density. These methods are sometimes considered to provide a good balance between accuracy and computational efficiency (Senn and Thiel, 2009). The last category includes the semi-empirical methods (e.g. AM1, PM3, MNDO), where empirical functions are introduced to simplify the most complex steps of *ab initio* calculations. This however compromises accuracy.

Due to computational complexity, electronic structure calculations are currently restricted to systems of a few hundred atoms. This is obviously insufficient to study an enzyme consisting of thousands of atoms. Enzymatic reactions are therefore more appropriately studied by the hybrid quantum mechanics/molecular mechanics (QM/MM) method, in which the system is separated into one "QM region" treated quantum mechanically and one "MM region" described by a classical force field (Warshel and Levitt, 1976; Gao and Truhlar, 2002; Senn and Thiel, 2009). The QM region is chosen such that it includes the atoms directly involved in the reaction, which comprises the substrate and active site residues. The potential energy for the nuclear coordinates is written as

$$E = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{QM/MM}} \tag{2.24}$$

where the three terms on the right-hand side denote the potential energy of respectively the QM region, MM region and the atoms constituting the boundary between the QM and MM regions. Special approaches need to be taken to model the latter (Gao and Truhlar, 2002; Senn and Thiel, 2009).

The potential energy surface obtained from the approximate solution of Equation (2.23) plays a role similar to the force field of Section 2.2.1. Each covalent modification that takes place in a chemical reaction pathway, is reflected in that the system crosses a high-energy barrier on the potential energy surface. In such a step, the system moves from a "reactant state", to a "product state", crossing a "transition state", which is a hyper surface separating the reactant state from the product state. One typically seeks to determine the minimum energy path, which is sketched in Figure 2.3. The reactant and product states correspond to local minima on the potential energy surface and are consequently determined by energy minimization with respect to the nuclear coordinates. The transition state on the minimum energy path is the configuration with the highest energy, which corresponds to a saddle point on the potential energy surface. Various methods exist for determining this configuration (Senn and Thiel, 2009).

Transition state theory (TST) (Laidler, 1987) relates the rate for crossing a poten-

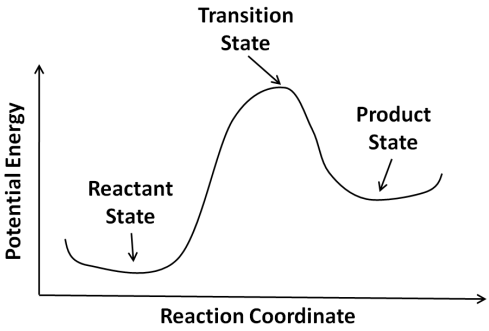


Figure 2.3: The minimum energy path of an arbitrary chemical reaction step. The reaction proceeds from left to right and the transition state needs to be formed before the transition is completed.

tial energy barrier to the Gibbs energy cost for forming the transition state, $\Delta G^{\ddagger}$, given that the system starts in the reactant state. TST approximates the rate by

$$k = \gamma(T)\frac{k_{\mathrm{B}}T}{h}\exp\left(-\frac{\Delta G^{\ddagger}}{k_{\mathrm{B}}T}\right) \tag{2.25}$$

where $\gamma(T)$ is a temperature-dependent transmission factor, which in basic TST is unity (cf. Equation (2.15)). In more sophisticated formulations, $\gamma(T) \neq 1$ in order to account for barrier re-crossing and quantum dynamical effects such as tunneling (Fernandez-Ramos *et al.*, 2007). In order to evaluate $\Delta G^{\ddagger}$, the free energy perturbation method is commonly applied (Senn and Thiel, 2009). This involves carrying out several MD or MC simulations of the system, where each simulation is run with the reaction coordinate fixed to a particular value, ϵ_i. These values are chosen such that the simulations span the path from the reactant to the product state. The free energy change between two adjacent reaction coordinate value, ϵ_i and ϵ_{i+1} is given by

$$\Delta G_{\epsilon_i,\epsilon_{i+1}} = k_{\mathrm{B}}T\ln\left\langle\exp\left(\frac{\Delta E_{\epsilon_i,\epsilon_{i+1}}}{k_{\mathrm{B}}T}\right)\right\rangle_{\epsilon_i} \tag{2.26}$$

where $\Delta E_{\epsilon_i,\epsilon_{i+1}}$ is the potential energy cost for progressing the reaction coordinate from ϵ_i to ϵ_{i+1}, and where $\langle\cdot\rangle_{\epsilon_i}$ denotes the ensemble average calculated in the ϵ_i simulation. Carrying out MD or MC simulations on a QM/MM potential energy surface is however very computationally demanding and is therefore usually done with a semi-empirical method describing the QM region.

Several protocols have been developed to combine the sampling efficiency of classical force fields or semi-empirical methods with the accuracy of DFT or *ab initio* methods. One example is the quantum mechanical thermodynamic cycle perturbation method by Rod and Ryde (2005a,b). This approach relies on carrying out several MD simulations with specific values of the reaction coordinate. These simulations, in which the entire QM region is kept fixed, use a classical force field for propagating the MM region. The obtained energies are then corrected using QM/MM with a DFT method describing the QM region. The method applied to

the methylation of catecholate catalyzed by catechol O-methyltransferase yielded an estimate of the free energy barrier which was in very good agreement with an experimentally determined value.

Application to Enzymes in Non-Aqueous Media Few quantum mechanical studies seem to have investigated the effects of non-aqueous solvents. One of the few studies was conducted by Colombo *et al.* (1999, 2000), who studied transesterification of vinyl acetate and *sec*-phenetyl, catalyzed by subtilisin. The solvents, which were modeled explicitly, were DMF, hexane and water. The authors aimed to rationalize the enantioselectivity of the enzyme. They did not determine the transition state explicitly, but used the tetrahedral intermediate state as model for it, which is reasonable according to the Hammond postulate (Fersht, 1999). The tetrahedral intermediate, which is illustrated in Figure 2.4, was first modeled by QM/MM using semi-empirical methods for the QM part. After minimizing the configurational energy, the charge distribution in the QM region was determined and used to assign partial charges to the individual atoms. The system was then simulated by MD using the AMBER force field, but with the charges determined by the QM/MM calculations. This procedure was repeated for both the S and R enantiomers of the substrate. The authors reported that the fast-reacting enantiomer (S) showed more favorable steric, hydrogen bonding and electrostatic interactions with the active site environment than the slow-reacting one (R). The groups of the substrate were solvated slightly differently in the different solvents. In particular, the phenyl group in the R enantiomer was in DMF and hexane oriented towards the solvent, due to the hydrophobic interactions. In water, the same phenyl group was oriented towards the hydrophobic pocket and was shielded from the solvent. Specific importance was attributed to that the charge distribution obtained from QM/MM was significantly different for the two enantiomers, especially near the substrate stereocenter. The choice of solvent was also observed to affect this charge distribution significantly.

The authors further carried out classical MD free energy perturbation calculations to determine the Gibbs energy difference between the tetrahedral intermediates of the R and S enantiomers to approximate the difference $\Delta\Delta G^{\ddagger}_{R,S}$ in free energy cost for forming the transition state for the two enantiomers. For DMF, the calculated $\Delta\Delta G^{\ddagger}_{R,S}$ was in good agreement with the experimental value. It was demonstrated to be crucial however, to use partial charges obtained from the QM/MM calculation taking the enzyme and solvent environment into account. With partial charges obtained from QM/MM minimization of the substrate in gas phase, a significantly different value of $\Delta\Delta G^{\ddagger}_{R,S}$ was obtained.

Corresponding results for water and hexane were not reported. The study however demonstrates that electronic structure phenomena may be of importance when rationalizing the effects of non-aqueous solvents.

Also notable is the study of Foresti *et al.* (2009), in which an experimental investigation of the enantioselective esterification of ibuprofen and ethanol catalyzed by CALB, was supplemented by semi-empirical QM calculations on the PM3 level. The authors explored the hypothesis that water or ethanol molecules can bind covalently at the active site, and act as inhibitors. QM/MM was not employed; instead energy minimizations were performed on a system consisting of the active site residues and the inhibitor candidate molecule, neglecting the influence of the remaining part of

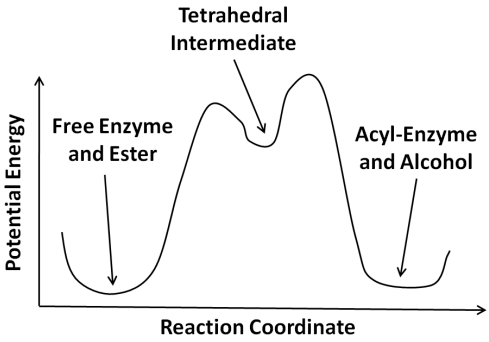


Figure 2.4: Minimum energy path for the acylation step of a serine hydrolase. The enzyme binds first the ester substrate covalently and forms the meta-stable tetrahedral intermediate. The alcohol part of the substrate is then released, which leaves the acyl group bonded to the enzyme.

the enzyme and the solvent environment. It was shown to be enthalpically feasible for both ethanol and water to bind the active site histidine of the free enzyme, which would result in inactivation of the enzyme. Transition states for forming these "dead-end complexes" were however not considered, and it is thus unclear if forming the complexes is kinetically feasible.

2.2.4 Summary

Enzymes in non-aqueous media have been studied by MD simulation by several researchers. General conclusions are that the enzyme structure is preserved but the flexibility is decreased as compared to in aqueous media. It has been shown that the organic solvent and the hydration level have a measurable effect on protein properties determined in simulation. The field could yet benefit from the development of a protocol for rigorous control of the hydration level, characterized by the water activity. This will allow more quantitative comparisons of simulations carried out in different solvents, and more quantitative comparisons of simulations and experiments performed at fixed water activities.

QM and QM/MM methods appear rather unexplored for studying enzymes in non-aqueous media. The results of Colombo *et al.* (1999, 2000) demonstrate however that they can supply useful information. What these methods can provide is information about the solvent effects on the free energy barriers for the covalent steps of catalysis. Exploring this is however beyond the scope of this work. Firstly, this is because transition-state stabilization is only one of several hypothesized ways solvent can affect the activity or selectivity of enzymes. Secondly, it is not necessarily covalent modifications that limit the reaction rate in enzyme catalysis. Conformational rearrangements could very well be more critical for this (Frauenfelder, 2008). A QM/MM study investigating medium effects on the free energy barriers would still be worthwhile, but if a thorough understanding on the medium effects on the conformational dynamics first is achieved, and if means to control the water activ-

ity are established, one would be in a better position to conduct and interpret the results of such a QM/MM investigation.

3

Molecular Dynamics Study of *Candida Antarctica* Lipase B - Part I

This chapter describes a molecular dynamics (MD) study of *Candida antarctica* lipase B (CALB) in pure water, acetone and hexane. For the two organic solvents, simulations were carried out with different amounts of water present. The purpose of this study is to investigate whether effects of solvent and hydration on CALB structure and dynamics can be observed in MD simulations. In this chapter, no attempts are made to rigorously measure or control the water activity. This is the focus of Chapter 7.

As apparent from the overview in the previous chapter, solvent and hydration effects on the flexibility of enzymes have been observed before. In particular, Trodler and Pleiss (2008) simulated CALB in water and five organic solvents, while Branco *et al.* (2009) studied CALB in a gaseous mixture of water and argon and analyzed flexibility at different water activities. No simulation study seems however to have investigated the effects of different hydration levels on CALB structure and dynamics in organic solvents. Considering the studies of *Fusarium solani* cutinase by Soares *et al.* (2003); Micaêlo *et al.* (2005) and Micaêlo and Soares (2007), one could expect that the hydration level does affect the flexibility in such environments. Cutinase and CALB have however different catalytic properties in non-aqueous media, which has been observed in experiments. While cutinase displays maximum esterification and transesterification activity when the thermodynamic water activity is between 0.2 and 0.7 (Vidinha *et al.*, 2003), the catalytic activity of CALB for these reactions is usually reported to be maximal at extremely low water activity (Humeau *et al.*, 1998; Chamouleau *et al.*, 2001; Petersson *et al.*, 2006; Mora-Pale *et al.*, 2007; Foresti *et al.*, 2009; Leonard-Nevers *et al.*, 2009). Simulations of CALB might therefore as well show different trends than those observed for cutinase and might provide further understanding on the molecular effects of enzyme hydration.

Another important purpose of the investigation in this chapter is that it provided the author with a deeper understanding of the craft of MD simulation of proteins and the challenges arising in the particular case of CALB. Hopefully, this chapter will serve a similar purpose for the reader.

3.1 CALB Structure and Function

CALB is an enzyme widely used in industrial applications (Anderson *et al.*, 1998). It is commonly employed as catalyst for esterification or transesterification, for instance

in the production of biodiesel (Su and Wei, 2008) or monoacyl glycerols (Damstrup *et al.*, 2005) mentioned in Chapter 1, and is known to remain active in both polar and non-polar organic solvents, even at very dry conditions.

In total, four crystal forms of CALB have been resolved. Two of them, one orthorhombic (pdb-ID: 1TCA) and one monoclinic (pdb-ID: 1TCB and 1TCC), were derived from the "wild-type" CALB, i.e. without any ligand molecule bound to the active site (Uppenberg *et al.*, 1994). The other two structures were obtained by co-crystallization of CALB either with the detergent Tween80 (pdb-ID: 1LBT) or with a phosphonate inhibitor (pdb-ID: 1LBS). The inhibitor was covalently bound to the active site serine residue, while the detergent was bound to the active site non-covalently.

CALB consists of 317 residues and belongs to the α/β-hydrolase family. The secondary structure includes a seven-stranded β-sheet, of which the first two strands are mutually anti-parallel, while the last six strands are parallel. These β-strands are alternated by ten α-helices and fifteen loop segments. The ten residues at the C-terminal form a β-hairpin. Figure 3.1 shows the 3D structure of CALB and Figure 3.2 shows the secondary structure topology.

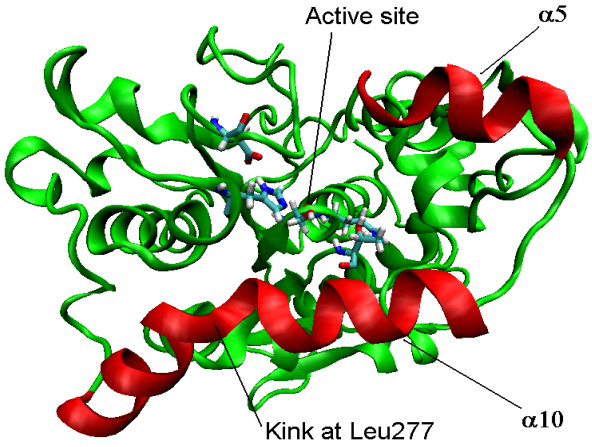


Figure 3.1: Image of CALB based on crystal structure (1TCA) (Uppenberg *et al.*, 1994), generated by VMD (Humphrey *et al.*, 1996). Proposed lid candidates, helices $\alpha 5$ and $\alpha 10$ are marked with red (Uppenberg *et al.*, 1994; Skjøt *et al.*, 2009), and the residues of the catalytic triad (Ser105, His224, Asp187) and the oxyanion hole (Thr40, Gln106) are indicated.

The active site consists of a catalytic triad including Ser105, His224 and Asp187 (Figure 3.3). During catalysis, the serine and histidine residues act as nucleophile and base, respectively, as the serine binds the substrate covalently while the serine H atom is temporarily transferred to the histidine. The aspartic acid residue is hydrogen bonded to the histidine side chain constraining its configuration relative to the serine. The oxyanion hole is formed by the backbone NH groups of the Thr40 and Gln106 and the side-chain OH group of the threonine residue. These groups are

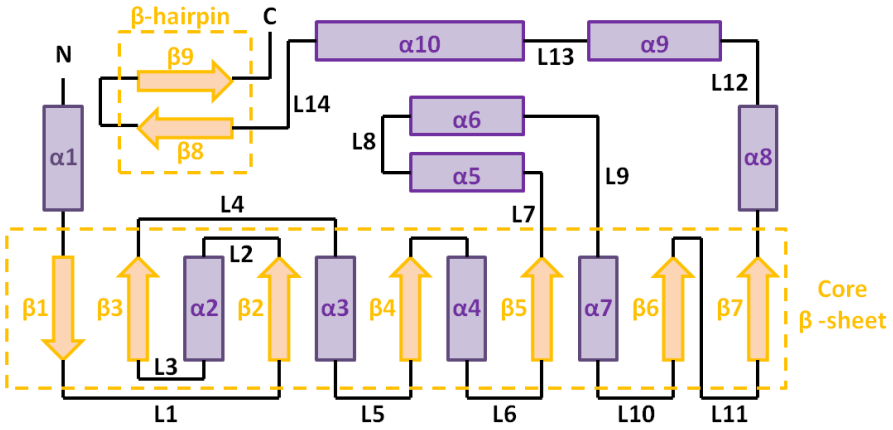


Figure 3.2: Secondary structure diagram for CALB (Uppenberg *et al.*, 1994). Notation for loop regions has also been introduced.

marked in Figure 3.3. The H atoms of these groups are directed towards the carbonyl O atom of the bound substrate, stabilizing the accumulated negative charge.

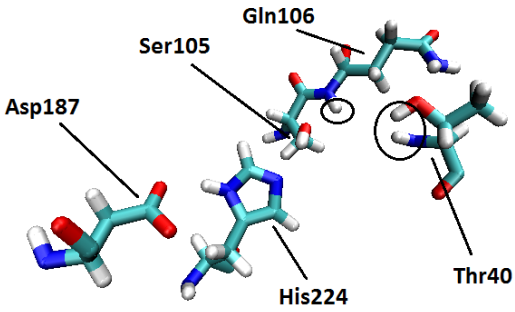


Figure 3.3: CALB active site region including the catalytic triad, Asp187, His224 and Ser105, and the oxyanion hole residues, Thr40 and Gln106. The hydrogen atoms of the oxyanion hole are marked with black circles.

Several lipases posses a lid, which typically consists of an α-helix connected to the rest of the protein via two loop segments, which allows the lid to "open" or "close" through a hinge-like motion. The active site is accessible in the open conformation, but blocked by the lid in the closed conformation. Commonly, the lipase assumes the closed conformation in aqueous environments. The lid opens when protein is located at the interface between a hydrophilic and a hydrophobic phase, a phenomenon termed interfacial activation (Dodson *et al.*, 1992; Derewenda, 1994). For CALB, a closed conformation has not yet been observed in any crystallographic study.

Table 3.1: Secondary structure elements of CALB with notation for α-helices and β-strands as given by Uppenberg *et al.* (1994). Notation for loop regions has been introduced for convenience.

Residue	Helix	Residue	β-strand	Residue	Loop
13–18	α1	20–22	β1	1–12	NT
44–57	α2	33–37	β2	23–32	L1
76–93	α3	62–66	β3	38–43	L2
106–117	α4	99–104	β4	58–61	L3
142–146	α5	125–131	β5	67–75	L4
152–156	α6	179–183	β6	94–98	L5
162–169	α7	208–211	β7	118–124	L6
212–216	α8	309–310	β8	132–141	L7
226–242	α9	313–314	β9	147–151	L8
268–287	α10			157–161	L9
				170–178	L10
				184–207	L11
				217–225	L12
				243–267	L13
				288–308	L14
				315–317	CT

The α-helices α5 and α10 were initially proposed as lid candidates, as they are located right at the rim of the active-site pocket, as shown in Figure 3.1. Especially interesting was α5 since in two of the resolved crystal structures (1TCC and 1LBS), the helix was "disordered", i.e. not having a well-defined conformation (Uppenberg *et al.*, 1994, 1995). Its role as lid was however ruled out since no experimental results seemed to support it. Martinelle *et al.* (1995) reported that CALB does not display interfacial activation. In their study, hydrolysis of p-nitrophenyl acetate in aqueous solution, was determined. With CALB as catalyst, the reaction rate dependence on substrate concentration obeyed the conventional Michaelis-Menten kinetics. This was seen also below the critical micelle concentration, where the substrate molecules are too few to form aggregates. This was compared to the same reaction, catalyzed by *Humigicola lanuginosa* lipase, which has a lid. This enzyme was inactive below the critical micelle concentration, but was activated when higher substrate concentrations were attained.

Recently, Skjøt *et al.* (2009) performed a mutational study of CALB in combination with MD simulations of CALB in pure water, in order to determine the role of α5. In the simulations, they observed that this helix unfolded, and the resulting loop moved towards the helix α10, thereby blocking the active site. In the (experimental) mutation study, the sequence of the region around α5 was exchanged for the corresponding sequence of homologous proteins. This resulted in mutants having either larger or smaller "lids", which had significant impact on the enzymatic activity and enantioselectivity. In the light of these results, the authors proposed that α5 indeed functions as a lid for CALB.

Trodler and Pleiss (2008) carried out MD simulations of CALB in water and five organic solvents and Branco *et al.* (2009) simulated CALB in a gaseous mixture of water and argon. In both those studies, the $\alpha 5$ region was seen to be very flexible, but no unfolding of this helix was reported. The flexibility of $\alpha 10$ was also reported to be rather high, although not as high as that of $\alpha 5$.

Given that the experimental support is dispersive, it is unclear whether CALB has a lid. The dynamics of the helices $\alpha 5$ and $\alpha 10$ nevertheless appear of central importance for the catalytic properties.

3.2 Simulation Procedure

MD simulations of CALB were carried out in acetone, hexane or pure water. In the acetone and hexane simulations, specific amounts of water molecules were included. The significance of this is discussed in Section 3.2.1. The procedure for generating the initial frame for each simulation is described in Section 3.2.2 along with the selection of force field. Section 3.2.3 finally gives the details of the MD simulations carried out.

3.2.1 How much Water?

It is well established that enzymes in non-aqueous media need to be hydrated to some extent in order to remain active (Zaks and Klibanov, 1988a). Simulations of proteins in organic solvents should thus include a certain number of water molecules. How many water molecules to include is however not obvious as the precise number of water molecules required for catalytic function is unknown. A common approach in literature is to include the "crystal water", i.e. those water molecules that are retained on the protein surface or in the protein interior during crystallization and are included in the pdb entry (Yang *et al.*, 2004; Trodler and Pleiss, 2008; Cruz *et al.*, 2009). Another approach is to include enough water to create a water monolayer covering the entire protein surface (Yang *et al.*, 2004). Though both approaches seem reasonable, it is not clear how one should select the number of water molecules, such that protein dynamics in different solvents can be compared.

For the investigation described in this chapter, several systems with different number of water molecules were simulated, in order to investigate their significance.

3.2.2 System Setup and Force Fields

The crystal structure coordinates of CALB were obtained from the protein data bank[1] (Berman *et al.*, 2000). As several structures are available, the best resolved one, 1TCA, which has a resolution of 1.55 Å (Uppenberg *et al.*, 1994) was used. The entry includes coordinates for CALB and 286 crystal water molecules. For simulations including fewer water molecules, those with the lowest B-factors were retained. For simulations including 286 or more water molecules, all crystal waters were retained, and additional water molecules were introduced using the software

[1] http://www.pdb.org/

Table 3.2: Summary of CALB simulations listing the number of water and organic solvent molecules, the total number of atoms and the number of replica simulations carried out with different starting velocities. Also listed is a short identifier for each simulation introduced for reference purposes.

Solvent	#water	#solvent	#atoms	#simulations	ID
Acetone	100	2200	26925	3	a100(a)–(c)
	286	2200	27483	5	a286(a)–(e)
	500	2200	28125	3	a500(a)–(c)
	1000	2200	29625	5	a1000(a)–(e)
	2400	1800	29825	3	a2400(a)–(c)
	4900	1300	32325	3	a4900(a)–(c)
Hexane	286	1150	28483	3	h286(a)–(c)
	1000	1150	30625	3	h1000(a)–(c)
Water	9500	-	33125	5	w(a)–(e)

VMD[2] in combination with the plug-in SOLVATE (Humphrey *et al.*, 1996). For the organic solvent simulations, the CALB-water complex was placed in a cubic box containing either acetone or hexane. The organic solvent molecule coordinates were taken from the last frame of an MD simulation of pure acetone or hexane of at least 500 ps. Solvent molecules closer than 2.5 Å to the CALB/water complex were removed. For the simulations of CALB in pure water, the entire simulation box was built using SOLVATE. The numbers of solvent molecules included in each simulation are listed in Table 3.2. It was ensured that enough solvent molecules were included in the simulation box, such that the protein did not interact with its periodic images.

The active site histidine (His224) was defined as neutral with the proton placed on N_δ, allowing for the essential hydrogen bond with Asp187. Asp134 was also defined as neutral, in accordance with the hypothesis of Uppenberg *et al.* (1994) that the neutral Asp134 side chain participates in a hydrogen bond network which ultimately promotes the electrophilic environment of the oxyanion hole. The pK_a of Asp134 was estimated to 7.25 using PROPKA 2.0 (Li *et al.*, 2005; Bas *et al.*, 2008), which furthermore supports the treatment of Asp134 as neutral. All remaining Arg, Asp, Glu and Lys residues were defined as ionized. In this treatment, the protein molecule was uncharged and the ionic strength was zero in all simulations.

The CHARMM27 force field (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) was used to model the protein, as well as the hexane molecules. For acetone molecules, the parameters reported by Martin and Biddy (2005) were used, while the TIP3P model with flexible bonds (CHARMM version) (MacKerell Jr. *et al.*, 1998) was employed for water.

[2]http://www.ks.uiuc.edu/Research/vmd/

3.2.3 Simulation Details

The simulations were carried out in the NPT ensemble (particle number, pressure and temperature constant) using the MD simulation program NAMD[3] (Phillips *et al.*, 2005). The velocity Verlet algorithm with a 1 fs step size was employed to integrate the equations of motion. Lennard-Jones forces were evaluated using a 12 Å cutoff and a 10 Å switching distance and non-bonded forces were evaluated using a pair list with an outer radius of 14 Å. Periodic boundary conditions were employed in the x, y and z directions, and electrostatic forces were evaluated using the particle mesh Ewald method with a grid spacing smaller than 1 Å. Temperature and pressure were maintained at 298.15 K (25 °C) and 1 atm, respectively, using the Langevin thermostat with a damping constant of 5 ps^{-1}, and the Langevin piston with a period of 200 fs and a decay constant of 500 fs. Coordinates were saved every 500 fs.

Prior to simulation, a 500 step conjugate gradient minimization of the configurational energy was carried out. During minimization, the C_α atoms were constrained to their crystal structure positions. The C_α atoms were constrained also during the first 300 ps of simulation, which were followed by a 200 ps run with the C_α atoms restrained by a harmonic potential. The force constant was set to 10, 5, 1 and 0.1 kcal/mol/Å^2 in 50 ps intervals. This procedure allowed the simulation box volume, solvent molecules and protein side chains to equilibrate before allowing protein backbone motion. This was followed by unconstrained simulation of approximately 10 ns, of which the final 6 ns was used for analysis.

3.3 Hydration Level

In the CALB simulations in acetone or hexane, the water molecules were initially located around the protein. In hexane, the water remained close to the protein surface throughout the simulation while in acetone, a significant amount of water left the surface and mixed with the bulk medium. The hydration level is here quantified by the number of water molecules in the first solvation shell of CALB, which here denotes the water molecules whose O atom is within 3.5 Å of any non-hydrogen CALB atom. This distance was chosen since radial distribution functions (RDFs) of water molecules around protein residues typically had the first minimum at 3.5 Å. This definition therefore ensures that the first hydration shell contains all water molecules included in the first peak of the RDFs It is noteworthy that 3.5 Å also was used in the studies of Schröder *et al.* (2006) and Branco *et al.* (2009).

Figures 3.4(a)–(d) compares the time evolution of the hydration level in the acetone, hexane and pure water simulations. In acetone, the hydration level decreased during the first 4 ns (Figures 3.4(a) and (b)), while it remained constant throughout the simulations in hexane (Figure 3.4(c)). This is expected since water is practically insoluble in hexane, while fully miscible with acetone. In pure water, the hydration level increased slightly during the first two ns (Figures3.4(d)), which probably was due to that the surface area of the protein increased slightly during the same period. This was confirmed by evaluation of solvent-accessible surface area (SASA) which

[3]http://www.ks.uiuc.edu/Research/namd/

is described in Section 3.4.3.

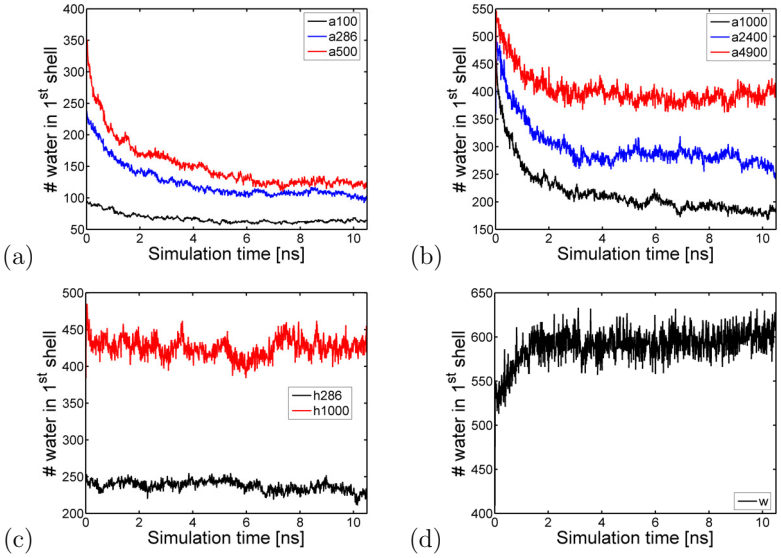


Figure 3.4: Number of water molecules in the first solvation shell vs. simulation time for representative CALB simulations in (a, b) acetone, (c) hexane and (d) water.

Comparing the simulations of CALB in acetone and hexane containing the same number of water molecules in total (286 or 1000), the average hydration level is at steady state significantly lower in acetone, as shown in Table 3.3. From the table, it is also apparent that depending on the medium, 75–95 % of the water molecules in the first shell were located near a hydrophilic residue. The lower the total hydration level was in acetone or hexane, the more the first hydration layer was concentrated around the hydrophilic residues at the CALB surface. The number of water molecules which were near a hydrophobic residue was roughly half the number of those near a hydrophilic one. This ratio seemed to be independent of the solvent. One should note that the hydration level of hydrophilic and hydrophobic residues should not necessarily sum up to the total hydration level since a single water molecule can be located near both a hydrophobic and a hydrophilic residue. Throughout this work, the notion of "hydrophobic" and "hydrophilic" residues follow the VMD selection convention (Humphrey *et al.*, 1996), counting Ala, Ile, Leu, Met, Phe, Pro, Trp and Val as hydrophobic and Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Lys, Ser, Thr and Tyr as hydrophilic.

For the sake of compatibility with other studies of proteins in non-aqueous media, the hydration level is also reported in terms of water weight / protein weight (w/w), employing the first shell definition.

In hexane, all the water molecules were throughout the simulation located around the protein surface. All of them were however not found within the first solvation shell. The second solvation shell was here defined to include the water molecules that were not in the first shell, but whose O atoms was located within 3.5 Å of

any O atom in the first shell. For the a100 simulations, the second shell was rather insignificant as it only contained two water molecules on the average, as seen in Table 3.3. The size of the second shell increased with increasing hydration. In h286, nearly all water molecules in the system were contained in the first and second hydration shell.

Table 3.3: Hydration level of CALB obtained from the simulations. The average number of water molecules in the first solvation shell around CALB, hydrophobic and hydrophilic residues (as defined in the text) is listed. The average number of water molecules in the second solvation shell, as defined in the text, is also shown.

System	#water in first shell	(w/w)	Hydrophilic	Hydrophobic	#water in second shell
a100	60 ± 1	3.3 %	56 ± 1	25.7 ± 0.1	2.3 ± 0.4
a286	107 ± 1	5.8 %	98 ± 2	41 ± 1	13 ± 1
a500	129 ± 1	7.1 %	116 ± 1	49 ± 1	28 ± 1
a1000	183 ± 2	10.0 %	162 ± 3	68 ± 1	58 ± 1
a2400	277 ± 3	15.1 %	235 ± 2	104 ± 1	162 ± 4
a4900	386 ± 3	21.1 %	316 ± 2	144 ± 2	349 ± 2
h286	240 ± 1	13.1 %	211 ± 2	95 ± 2	43 ± 1
h1000	433 ± 2	23.6 %	364 ± 3	155 ± 0.3	339 ± 2
w	595 ± 3	32.5 %	454 ± 3	235 ± 1	832 ± 5

3.4 Structure

3.4.1 Root-Mean Square Deviation

The structural variation of CALB in each of the 33 simulations was first assessed by monitoring the root-mean square deviation (RMSD) δ. For a selected set of N atoms with positions $\boldsymbol{r}_1(t), \ldots, \boldsymbol{r}_N(t)$ at time t, the RMSD is measured from a reference structure in which the atoms are located at $\boldsymbol{r}_{(\mathrm{ref}),1}, \ldots, \boldsymbol{r}_{(\mathrm{ref}),N}$, according to

$$\delta(t) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{r}_i(t) - \boldsymbol{r}_{(\mathrm{ref}),i} \right)^2 \right)^{1/2} \tag{3.1}$$

Prior to the evaluation, the atoms are at each frame aligned to the reference structure, in order to remove translation and rotation of the entire selection. The RMSD for C_α atoms was evaluated for each simulation using the crystal structure coordinates (1TCA) as reference. In 16 of the 33 simulations (1 in pure water, 12 in acetone and 3 in hexane), the RMSD initially increased, but reached a plateau after approximately 1 ns. The RMSD remained near this plateau for the remaining simulation time. In the remaining 17 simulations, there was a drift in RMSD, which however in each case could be attributed to one or several flexible regions. For each simulation, the RMSD curve did reach a plateau if the appropriate region(s)

was omitted. Depending on the simulation, these regions included the N-terminal (residues 1–10), the loop L1 (residues 23–32), the helix α5 and adjacent loop segment (residues 138–152), a part of the loop L11 (residues 190–202), the loop L13 and the adjacent helix α10 (residues 243–292), and the C-terminal (residues 308–317). Selected RMSD plots obtained from simulations carried out in acetone are shown in Figure 3.5. A larger selection of RMSD plots are shown in Appendix B.

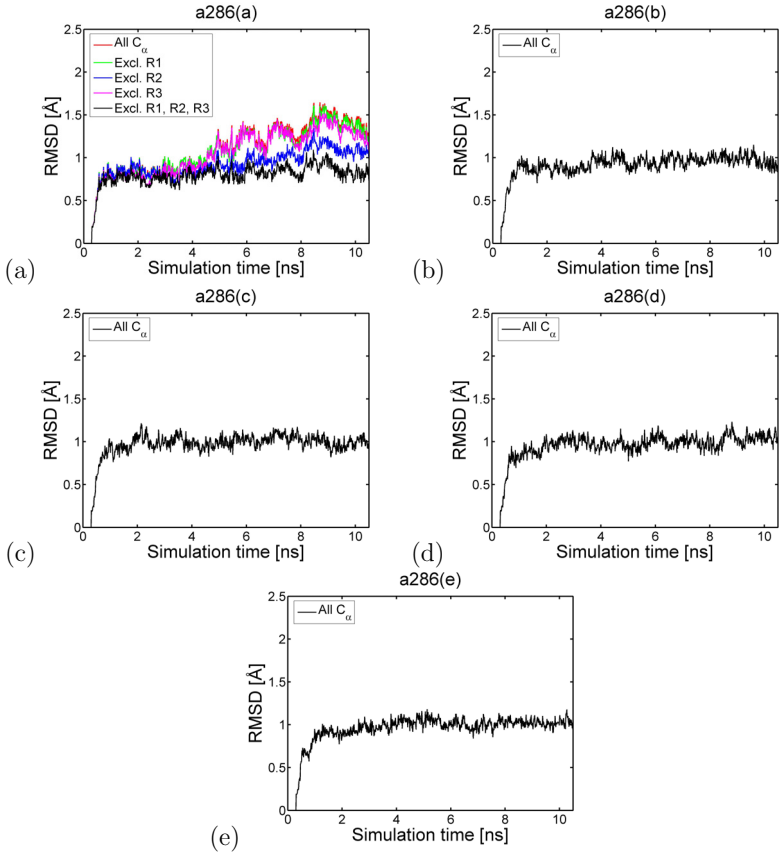


Figure 3.5: RMSD plots for CALB simulations a286(a)-(e) in acetone. Regions that need to be omitted from the calculation for obtaining stable RMSD in (a) comprise residues 1–10 (R1), 23–32 (R2), and 138–152 (R3).

The average RMSD over the last 6 ns was between 0.9 and 1.5 Å with all C_α atoms considered. Two simulations in pure water, w(b) and w(c), showed RMSD values of 1.8 and 2.2 Å, respectively. This could however be ascribed to fluctuations in the region around helix α5 and the N-terminal. Since fluctuations of the N- and C-terminals frequently caused a drift in the total RMSD, they were consistently omitted in the following structural analysis.

In order to detect local structural changes, the RMSD contribution from every individual C_α atom was calculated and averaged over the final 6 ns of simulation.

The CALB structure showed the largest deviation from the crystal structure in pure water, followed by hexane and acetone. Residues in three particular regions of CALB had in several simulations an average RMSD significantly exceeding 2 Å, namely α5 (water, acetone and hexane), L13 (hexane) and α10 (hexane) (see Figure 3.6). Interestingly, the helix α10 and the loop L13 which connects α10 with the helix α9, underwent structural changes in some of the hexane simulations, but neither in water or acetone.

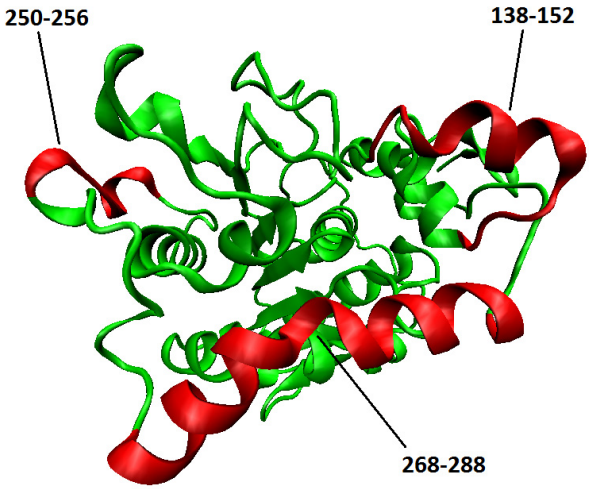


Figure 3.6: Image of CALB indicating the three regions displaying minor structural changes. The image was generated by VMD (Humphrey *et al.*, 1996).

The average RMSD of these three regions, as well as the overall RMSD is given in Table 3.4. The overall RMSD was apparently rather insensitive to the medium. It was slightly higher in pure water than in the organic solvents. For the acetone and hexane simulations, the overall RMSD displayed no significant dependence on the hydration level. The RMSD of the α5 region increased with increasing hydration in both acetone and hexane and the pure water simulations yielded the highest RMSD of this region. For the L13 region, the trend was reversed, as in acetone and hexane, the RMSD decreases with increasing hydration. The value obtained in pure water was significantly lower than the ones obtained in a100 and h286. For the α10 region, the RMSD values obtained in acetone appeared uncorrelated with hydration and were comparable to the values obtained in pure water.

The observed structural changes of the helices α5 and α10 are interesting since these regions are located at the rim of the active site pocket, and their importance for CALB function has been discussed in the literature (Uppenberg *et al.*, 1994, 1995; Martinelle *et al.*, 1995; Skjøt *et al.*, 2009). A more detailed analysis of the structural changes occurring in these regions is therefore given in the next section.

The loop region L13 appeared to be stabilized by water. The region contains a buried tyrosine residue (253), an exposed isoleucine residue (255) and a semi-exposed

aspartic acid residue (252). Possibly, the structural change that occurred upon dehydration was due to a tendency for the charged Asp252 to hide from the solvent, and for the hydrophobic Tyr253 and Ile255 to become more exposed. This was supported by the changes in solvent-accessible surface area (SASA) which measures the area of the part of the residue that is exposed to the solvent. Details of the SASA calculation are given in Section 3.4.3. The average SASA of the hydrophobic residues Tyr253 and Ile255 was respectively 58 Å^2 and 18 Å^2 higher in h286 than in the pure water simulations. For the charged Asp252, the average SASA was 18 Å^2 lower in h286 than in pure water. Going from pure water to a100, the SASA of Tyr253, Ile255 and Asp252 changed by respectively 13 Å^2, 18 Å^2 and -22 Å^2.

Table 3.4: Contributions to the total RMSD for selected CALB regions. Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

RMSD [Å]	All (11–307)	α5 (138–152)	L13 (250–256)	α10 (268–287)
a100	1.04 ± 0.03	2.7 ± 0.4	2.0 ± 0.6	1.0 ± 0.1
a286	1.07 ± 0.07	2.5 ± 0.1	1.6 ± 0.2	1.38 ± 0.06
a500	1.05 ± 0.09	3.0 ± 0.4	1.8 ± 0.6	1.1 ± 0.1
a1000	1.2 ± 0.1	3.2 ± 0.3	2.08 ± 0.3	1.4 ± 0.2
a2400	1.17 ± 0.04	3.6 ± 0.1	1.3 ± 0.4	1.4 ± 0.2
a4900	1.10 ± 0.02	3.3 ± 0.1	0.87 ± 0.07	1.1 ± 0.1
h286	1.2 ± 0.2	2.5 ± 0.2	2.3 ± 0.7	2.2 ± 1.0
h1000	1.18 ± 0.09	3.1 ± 0.1	1.1 ± 0.2	1.2 ± 0.3
w	1.33 ± 0.07	4.3 ± 0.4	1.2 ± 0.1	1.2 ± 0.1

3.4.2 Conformational Change of Helices α5 and α10

An interesting event occurring in several of the simulations was that the helix α5 (residues 142–146) was displaced with respect to the crystal structure and in some simulations unfolded, partially or completely. In the different simulations, the structure of α5 changed in different ways. Five qualitatively different terminal structures were identified. These structures, respectively denoted A, B, C, D and E, are shown in Figures 3.7(b)–(f) and described in detail below.

In the crystal structure (1TCA), the helix is defined by backbone hydrogen bonds between the CO groups of Ala141, Gly142 and Pro143 and the NH groups of Asp145, Ala146 and Leu147, respectively. The position of the helix is further constrained by side-chain hydrogen bonds between the side chain OH groups of Ser150 and Thr158 and the COO^- group of Asp145, which is pointing "inwards", towards the helix α6 (Figure 3.7(a)).

Backbone Hydrogen Bonds In all five pure water simulations, the 141–145 backbone hydrogen bonds was broken, usually within the first nanosecond of simulation, which resulted in an "unwinding" of the helix from the N-terminal direction.

The 142–146 hydrogen bond was broken in two of the simulations and the 143–147 bond in three. These bonds were usually maintained for several ns of simulation as a rule and were broken after the 141–145 bond.

The helix appeared as a whole more stable in the organic solvent simulations, especially at low hydration. The 141–145 bond was broken in all acetone simulations. The 143–147 bond was consistently maintained in a100 and a286, while broken in a4900. In a500, a1000 and a2400, the hydrogen bond was depending on the simulation either maintained or broken. The 142–146 bond was maintained in all acetone simulations except a1000(a)–(b) and a4900(b)–(c).

In the three hexane simulations h286(a)–(c), the three hydrogen bonds and the helical structure were well maintained throughout the simulations. In h1000(a)–(c), the 141–145 bonds was broken, while the other two bonds were retained.

The relatively low stability of the 141–145 bond is possibly related to the fact that the residue 143 is a proline residue, which due to its restricted Ramachandran ϕ-angle often is a helix breaker (Brändén and Tooze, 1999).

Displacement of Asp145 In all simulations, at least one of the two hydrogen bonds connecting Asp145 to Ser150 and Thr158 was broken. In simulations where $\alpha 5$ underwent a significant structural change, Asp145 was typically the residue undergoing the largest displacement, with respect to the starting structure. The displacement of Asp145 resulted in five qualitatively different situations. Which behavior occurred depended on the particular simulation, as indicated in Table 3.5.

In the first situation (A), both crystal structure hydrogen bonds of Asp145 were broken and the residue was oriented towards Lys308 and Arg309 (Figure 3.7(b)). Simultaneously, the two positively charged residues oriented themselves towards Asp145. The side chains did however not get close enough to form salt bridges. This situation occurred exclusively in hexane at low hydration.

In the second situation (B), the Asp145 COO^- group approached Arg309 of the C-terminal and formed a salt bridge (Figure 3.7(c)). This was primarily a result of side chain re-orientation, as the backbone atoms of the two residues did not undergo any significant displacement. This situation was quite exceptional, and occurred only in two of the acetone simulation.

In the third situation (C), the hydrogen bond between Asp145 and Thr158 was well maintained. The O^- of Asp145 not participating in the hydrogen bond was either stabilized by the backbone NH groups of Gly143, Leu144 and Asp145 or was exposed to the solvent (Figure 3.7(d)). This situation was encountered in acetone at low hydration.

In the fourth situation (D), the hydrogen bond between Asp145 and Ser150 was maintained. The Asp residue typically approached the positively charged side chain of Lys290 (Figure 3.7(e)). The residues got in cases sufficiently close to form a salt bridge. This situation occurred in pure water and in acetone and hexane at high hydration levels.

In the last situation (E), the C_α displacement of Asp145 was especially large, and the unfolded helix formed a loop extending out from the protein with the Asp145 COO^- group directed out into the water (Figure 3.7(f)). This situation, which corresponded to the highest degree of $\alpha 5$ unfolding, occurred in pure water and in acetone at high hydration levels. This situation also resulted in the largest C_α

displacement and solvent exposure of Asp145, as well as the largest structural change of the helix.

Table 3.5: The different types of behavior of Asp145, which are described in Section 3.4.2. The systems showing each type of behavior are listed along with the average C_α displacement and SASA of Asp145. The average RMSD of the α5 region for each situation is also shown. Standard error estimates were based on values from different simulations with similar behavior.

Type	Systems	Asp145 Disp. [Å]	Asp145 SASA [Å^2]	α5 (138–152) RMSD [Å]
A	h286	3.9 ± 0.2	71 ± 9	2.5 ± 0.3
B	a286, a1000	3.6 ± 0.8	56 ± 8	2.5 ± 0.4
C	a100, a286, a500	2.3 ± 0.2	39 ± 6	2.6 ± 0.1
D	a1000, a2400, a4900, h1000, w	3.7 ± 0.2	48 ± 4	3.3 ± 0.1
E	a500, a1000, w	8.4 ± 1.1	137 ± 7	4.2 ± 0.3

Consistency with Previous Results The observation that the helix α5 in water unfolds to become a loop region of high flexibility is consistent with the observation of Skjøt *et al.* (2009). They however observed that Pro143 moved towards the helix α10, "closing" for the entrance to the active site. This did not occur in the simulations described here, and by visual inspection of the terminal frame, the active site appears to be accessible in all simulations. Trodler and Pleiss (2008) and Branco *et al.* (2009) did not report any unfolding of α5.

As briefly mentioned in Section 3.1, the helix α5 is seen to have a well-defined helical structure in two of the x-ray resolved crystal structures, while it is seen to be disordered in two other structures (Uppenberg *et al.*, 1994, 1995). In two of the cases where the helical structure is observed, a hydrophobic molecule is located at the entrance to the active site, in contact with the side chain of Leu140. In 1TCB, it is a β-octyl glucoside detergent molecule, while it is the Tween80 detergent molecule in 1LBT. In cases where helix α5 is disordered (1TCC and 1LBS), no such molecule is present. In the orthorhombic crystal structure, 1TCA, α5 is helical despite the absence of a detergent molecule in the channel. Uppenberg *et al.* (1994) however pointed out that α5 may be stabilized by crystal packing. The side chain of Leu199 of a neighboring CALB molecule points into the active site and could here play the role of the detergent molecule stabilizing the helix α5.

The presence or lack of helical structure of α5 is reflected in the B-factors derived from the crystal structure data. When α5 is disordered, the C_α B-factors of residues 140–150 are significantly higher than the B-factors of remaining residues, as shown in Figure 3.8.

McCabe *et al.* (2005) determined the secondary structure content of CALB in hexane, toluene, 1,4-butanediol and water at neutral pH using circular dichroism. They reported that the α-helix content was insensitive to the solvent. A complete unfolding of α5 does however correspond to a change of the α-helix content of CALB

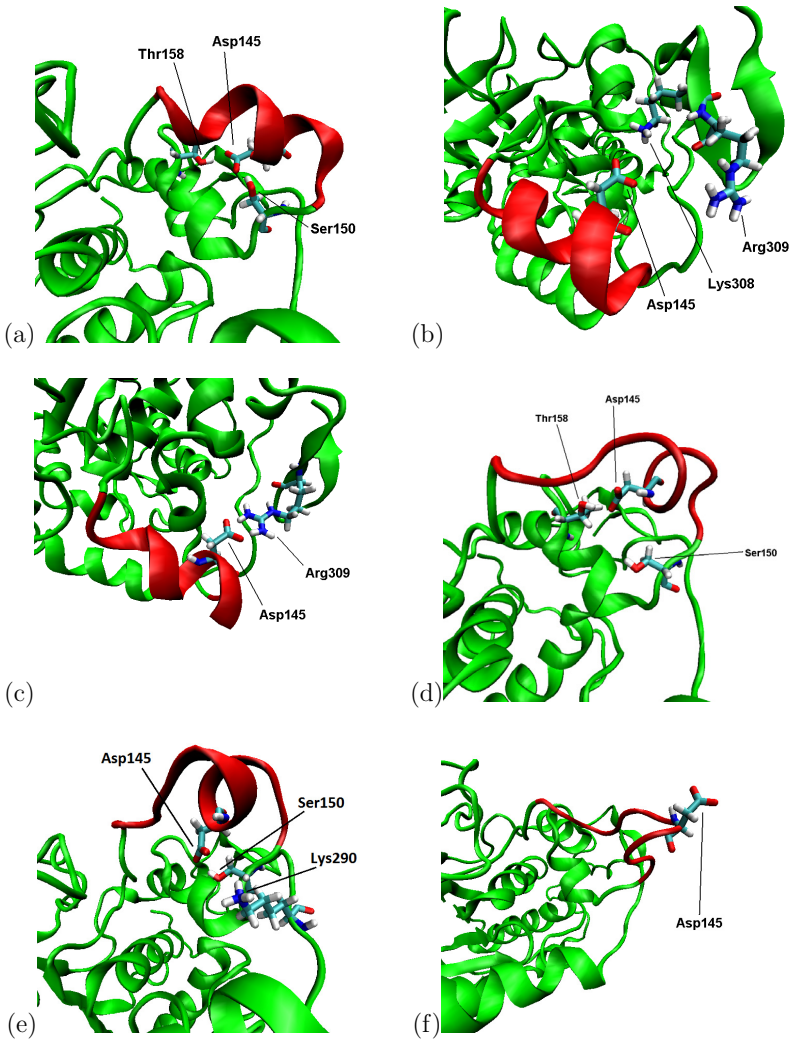


Figure 3.7: CALB conformations from (a) the 1TCA crystal and from the terminal frames of simulations representative of situation (b) A, (c) B, (d) C, (e), D and (f) E, as described in Section 3.4.2. Table 3.5 lists which simulations showed which type of behavior. In all figures, the region composed of residues 140–149 including the helix $\alpha 5$ is marked in red color and Asp145 is shown in "liquorice". Other residues shown are Ser150 (a, d, e), Thr158 (a, d), Lys290 (e), Lys308 (b), and Arg309 (b, c). The images were generated using VMD (Humphrey *et al.*, 1996).

by only 2 percentage points. This is well within the variation of the data presented by McCabe *et al.* (2005).

Considering the factors discussed above, the unfolding of $\alpha 5$ in the pure water simulations is not unexpected, and does not necessarily contradict the experimental findings.

Conformational Change of $\alpha 10$ In one of the three hexane simulations (h286(a)), the helix $\alpha 10$ partially unfolded. The unfolding occurred in the N-terminal part of the helix, right before the kink at Leu277 (see Figure 3.1). The event took place after 5 ns of simulation and comprised breaking of the backbone hydrogen bonds between Gln270/Ala274, Lys271/Ala275 and Val272/Ala276. The C-terminal part of the helix, which is part of the walls of the active site pocket, appeared to be unaffected. The event was not observed in any of the other simulations.

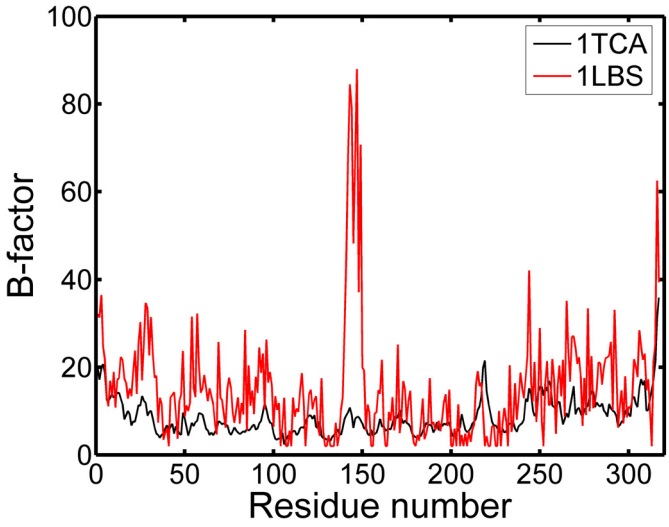


Figure 3.8: B-factors for C_α atoms derived from the crystal structures 1TCA (black) (Uppenberg *et al.*, 1994) and 1LBS (red) (Uppenberg *et al.*, 1995).

3.4.3 Solvent-Accessible Surface Area

Solvent accessible surface area (SASA) measures the outer surface area of a molecule. It is essentially the area traced by a spherical solvent particle of radius r_S which is rolled over the entire molecular surface. The SASA of CALB was evaluated using VMD (Humphrey *et al.*, 1996), which employs an algorithm which circumscribes each protein atom i with a sphere of radius $r_S + d_i$, where d_i is the van der Waals radius of the atom. A number of grid points are placed on random locations uniformly distributed over each such sphere. All grid points that lie inside a sphere circumscribed around another atom are removed. The remaining grid points span

the solvent accessible surface of the protein, and from these points, the area is deduced. Here, the average SASA of CALB was evaluated based on the final 6 ns of each simulation. A ball radius of $r_S = 1.4$ Å was used, corresponding roughly to the radius of a water molecule. Only non-hydrogen protein atoms were considered in the calculation. The average total SASA for all residues, hydrophilic residues and hydrophobic residues (defined in Section 3.3) are listed in Table 3.6. The average SASA contributions from each individual residue were also evaluated. Those values are not reported here, but selected values are employed in the discussion in Sections 3.4.1 and 3.4.2.

Table 3.6: Average SASA of CALB, determined from the simulations. Reported SASA values were respectively evaluated all residues, hydrophilic residues and hydrophobic residues (as defined in Section 3.3). Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

SASA [Å^2]	Total	Hydrophilic	Hydrophobic
a100	13334 ± 4	7480 ± 20	5860 ± 20
a286	13530 ± 90	7580 ± 40	5940 ± 60
a500	13520 ± 40	7580 ± 30	5940 ± 60
a1000	13700 ± 100	7670 ± 60	6010 ± 60
a2400	13810 ± 20	7740 ± 20	6070 ± 40
a4900	13900 ± 100	7810 ± 50	6030 ± 70
h286	13310 ± 90	7390 ± 20	5920 ± 80
h1000	14000 ± 70	7820 ± 30	6180 ± 40
w	14200 ± 100	8060 ± 80	6120 ± 40
Crystal (1TCA)	12130	6980	5150

The total SASA of CALB was largest in the pure water simulations. In the acetone simulations, the SASA increased with increasing hydration, as shown in Figure 3.9. The same trend was as well observed in hexane. The SASA of hydrophilic residues followed the same trend in both solvents. This was expected, since water should stabilize charged and polar residues which are exposed to solvent. Perhaps less expected, the SASA of hydrophobic residues also increased with increasing hydration in acetone and hexane. The SASA of these residues was however much less sensitive as compared to the SASA of the hydrophilic ones. Probably, the increase in SASA upon increased hydration was driven by the tendency for hydrophilic residues to become more exposed. The hydrophobic residues, being less sensitive to the medium, adopted the conformation that best stabilized the hydrophilic residues.

Figure 3.9 shows also that the total SASA was lower in hexane than in acetone, if values are compared at similar hydration levels. This is reasonable as hydrophilic residues should be better stabilized by acetone than hexane molecules.

Trodler and Pleiss (2008) observed that the total SASA of CALB was highest in water, and decreased with the hydrophobicity of the organic solvent. This is consistent with the results presented here.

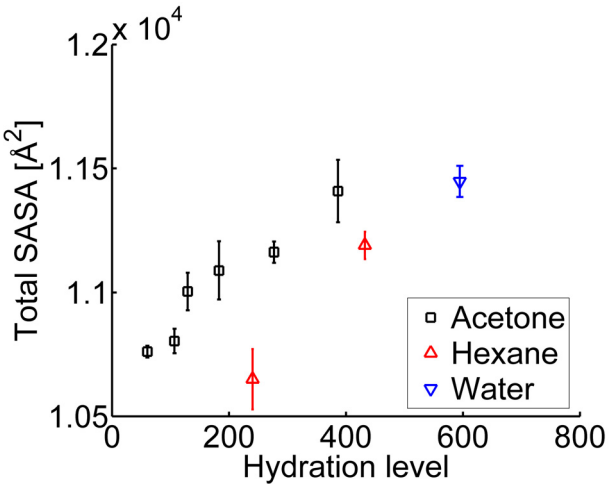


Figure 3.9: Average total SASA evaluated from the simulations carried out in acetone (black square), hexane (red triangles) and pure water (blue triangles). Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities. Hydration level refers to the number of water molecules in the first solvation shell (see Section 3.3 for details).

3.5 Flexibility

Protein flexibility was characterized by the B-factors, which for an atom denoted i are defined by

$$\beta_i = \frac{8\pi^2}{3} \left\langle |\boldsymbol{r}_i - \langle \boldsymbol{r}_i \rangle|^2 \right\rangle \tag{3.2}$$

where $\boldsymbol{r}_i$ denotes the position of the atom and $\langle \cdot \rangle$ denotes time average. As with the RMSD calculation, the protein structure trajectory was aligned to the crystal structure (1TCA) prior to the calculation, in order to remove translational and rotational motion of the entire protein. The average structure was calculated from the last 6 ns of the trajectory and the B-factors were obtained from the C_α atom fluctuations around the computed average structure. The B-factors obtained for the nine studied systems are shown in Figures 3.10–3.13. These were in good qualitative agreement with the crystal structure B-factors of Figure 3.8 and were furthermore overall consistent with the B-factors from simulations reported by Trodler and Pleiss (2008) and Skjøt *et al.* (2009).

The highest local flexibility was observed in the pure water simulations for the N-terminal (residues 1–20) (not shown in the figures), the region around the helix $\alpha 5$ (138–152) and the C-terminal (308–317) (not shown in the figures). The N-terminal was also rather flexible in hexane. The $\alpha 5$ region exhibited also high flexibility in hexane and in acetone. The high flexibility of the termini was in some simulations due to slow, transient motion, which sometimes is observed in protein simulations (see also discussion of Section 3.4.1). For this reason, the termini were omitted from the analysis. In a few simulations, similar slow transient motion was seen in the

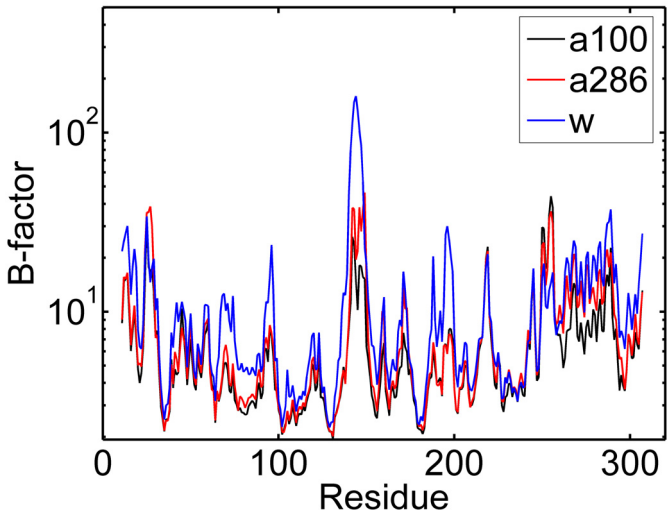


Figure 3.10: B-factors for C_α atoms obtained from the simulations a100 (black), a286 (red) and w (blue). Each curve is the average over 3–5 replica simulations (see Table 3.2), and the N- and C-terminals are omitted.

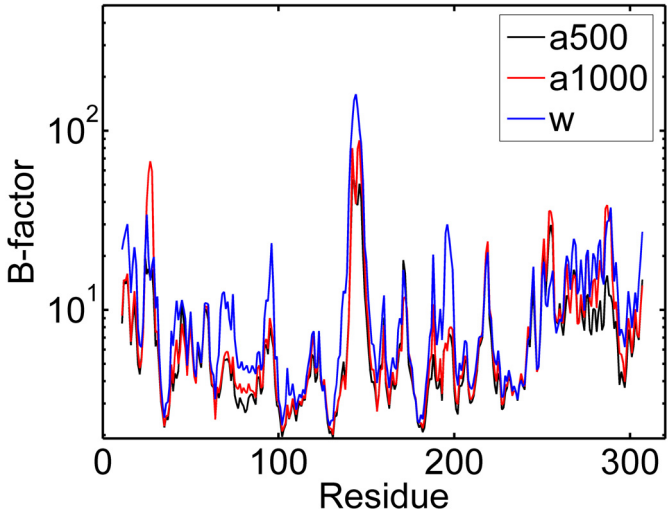


Figure 3.11: B-factors for C_α atoms obtained from the simulations a500 (black), a1000 (red) and w (blue). Each curve is the average over 3–5 replica simulations (see Table 3.2), and the N- and C-terminals are omitted.

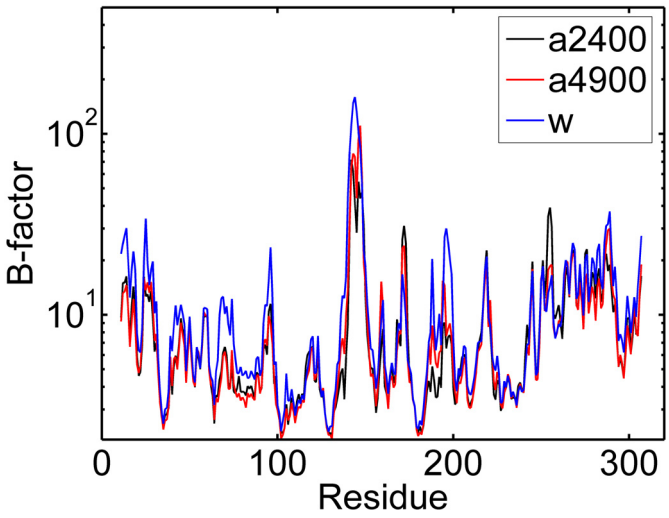


Figure 3.12: B-factors for C_α atoms obtained from the simulations a2400 (black), a4900 (red) and w (blue). Each curve is the average over 3–5 replica simulations (see Table 3.2), and the N- and C-terminals are omitted.

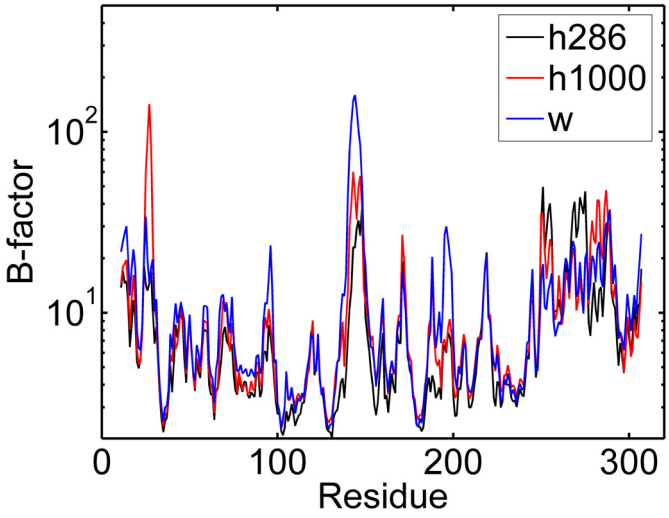


Figure 3.13: B-factors for C_α atoms obtained from the simulations h286 (black), h1000 (red) and w (blue). Each curve is the average over 3–5 replica simulations (see Table 3.2), and the N- and C-terminals are omitted.

loop L1 (23–32), which resulted in a sharp peak in Figures 3.10–3.13. For these simulations, L1 was as well omitted from the quantitative analysis.

The loop sections composed of residues 67–75 (L4), 94–98 (L5) and 195–199 (L11), were significantly more flexible in pure water than in the presence of organic solvent, as shown in Figures 3.10–3.13. These segments showed low flexibility in acetone and hexane, regardless of the hydration level.

The B-factor averaged over all CALB residues is shown in Table 3.7 and Figure 3.14. The overall flexibility was significantly higher in pure water than in the organic solvents. In both acetone and hexane, the flexibility increased with increasing hydration. At approximately similar hydration levels, the flexibility in acetone did not differ significantly from that in hexane (Figure 3.14).

In the regions where structural changes were observed (see Section 3.4.1 and Figure 3.6), namely α5 (138–152), L13 (250–256) and α10 (268–287), the flexibility was also high for all studied systems, as indicated in Figures 3.10–3.13. The flexibility of these regions was sensitive to the solvent. The average B-factors of those regions are reported in Table 3.7, and Figures 3.15–3.17.

For α5, the flexibility increased with increasing hydration in both acetone and hexane, and was highest in pure water. At approximately similar hydration levels, the flexibility was however significantly lower in hexane than in acetone (Figure 3.15). It was suggested in Section 3.4.2 that the unfolding of α5 was driven by the solvent-exposure of the negatively charged Asp145. As this residue is more favorably solvated by acetone than hexane, acetone could be expected to induce a higher flexibility than hexane.

The loop section L13 showed the opposite trend. The flexibility decreased with increasing hydration and was lowest in pure water. At approximately similar hydration levels, the flexibility appeared furthermore higher in hexane than in acetone, although the statistical uncertainties were relatively large (Figure 3.16). As discussed in Section 3.4.1, the structural change of L13 comprised the exposure of the hydrophobic Tyr253. This residue should be more favorably solvated by hexane than acetone.

For α10, there was no apparent correlation between hydration level and flexibility.

The regions of solvent-dependent flexibility identified here corresponded fairly well to those identified by Trodler and Pleiss (2008). They however observed also the flexibility of the loop L12 (215–222) to be solvent-dependent. In the present study, the flexibility of L12 was almost exactly the same in all studied systems. The increase in flexibility of α5 upon an increased hydration level was also observed in the study of Branco *et al.* (2009). They furthermore reported that the flexibility of L13 decreased with increasing hydration, consistent with the present findings.

The lower protein flexibility observed in organic media as compared to in water is consistent with several previous simulation studies (Norin *et al.*, 1994; Toba and Merz, 1996; Zheng and Ornstein, 1996a,c; Soares *et al.*, 2003; Trodler and Pleiss, 2008). In these studies, the phenomenon was attributed to the lower capability of the organic solvent to participate in hydrogen bonds and to shield protein-protein electrostatic interactions. The increase in flexibility with increasing hydration is consistent with the hypothesis that the water layer at the protein surface acts as a lubricant, promoting flexibility (Broos *et al.*, 1995).

Trodler and Pleiss (2008) proposed that certain "slowly exchanged" water mole-

Table 3.7: B-factors obtained from the simulations for C_α atoms averaged over certain CALB regions. Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities. The crystal structure B-factors (Uppenberg *et al.*, 1994) of corresponding regions are also listed. Slowly exchanged water refers to individual water molecules having a B-factor less than 25 Å^2 in the particular simulation. For simulations showing transient motion of L1, this loop was omitted in calculating the average.

Average B-factor [Å^2]	a100	a286	a500	a1000	a2400	a4900
11–307* (All)	6.6 ± 0.2	7.7 ± 0.5	7.3 ± 0.3	8.7 ± 0.8	8.7 ± 0.3	9.1 ± 0.3
138–152 (α5)	13 ± 2	22 ± 5	25 ± 5	34 ± 9	30 ± 4	42 ± 10
250–256 (L13)	30 ± 3	26 ± 3	21 ± 2	26 ± 3	24 ± 2	16 ± 1
268–287 (α10)	9.7 ± 0.1	15 ± 1	10.6 ± 0.3	17 ± 3	16 ± 2	15 ± 2
# Slowly exchanged water	20 ± 1	18.0 ± 0.4	17.3 ± 0.9	16.4 ± 0.7	14 ± 1	11 ± 2

Average B-factor [Å^2]	h286	h1000	w	1TCA
11–307* (All)	8.6 ± 0.9	9.8 ± 0.9	12.3 ± 0.9	7.9
138–152 (α5)	17 ± 5.0	28 ± 4	70 ± 10	7.8
250–256 (L13)	30 ± 10	25 ± 3	14.2 ± 0.5	15.4
268–287 (α10)	23 ± 9	24 ± 7	19.2 ± 0.8	11.3
# Slowly exchanged water	19 ± 1	11.0 ± 0.6	11.0 ± 0.9	

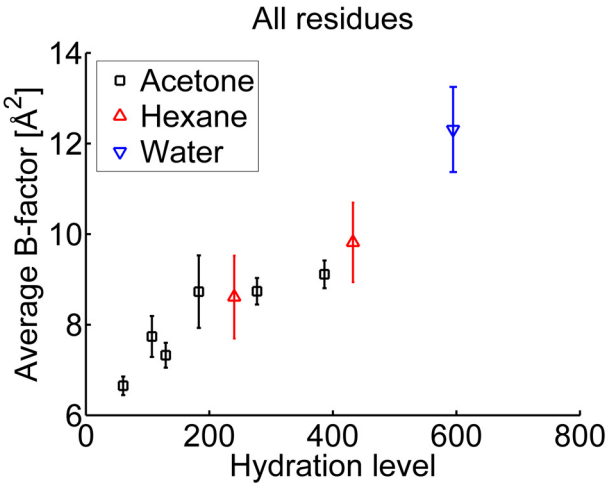


Figure 3.14: B-factor averaged over residues 21–307, as obtained from simulations in acetone (black squares), hexane (red triangles) and pure water (blue triangles). Residues 23–32 were omitted if loop L1 showed transient motion. Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

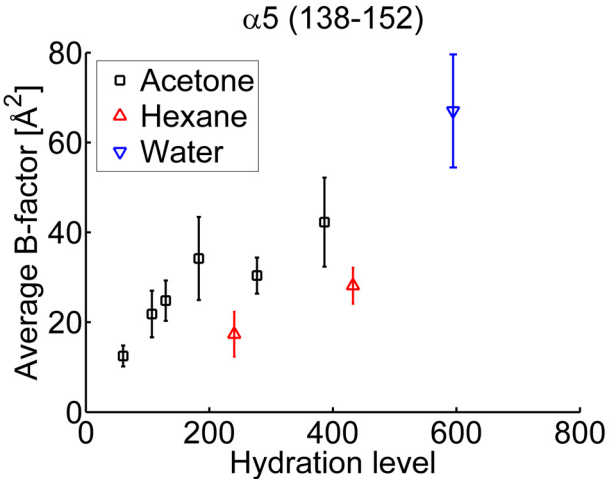


Figure 3.15: B-factor averaged over residues 138–152, as obtained from simulations in acetone (black squares), hexane (red triangles) and pure water (blue triangles). Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

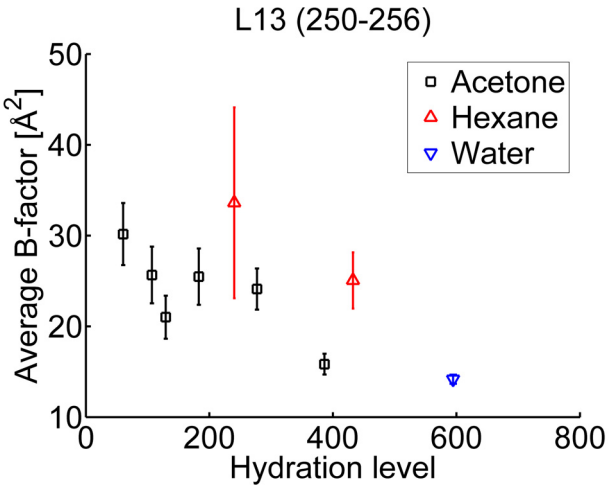


Figure 3.16: B-factor averaged over residues 250–252, as obtained from simulations in acetone (black squares), hexane (red triangles) and pure water (blue triangles). Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

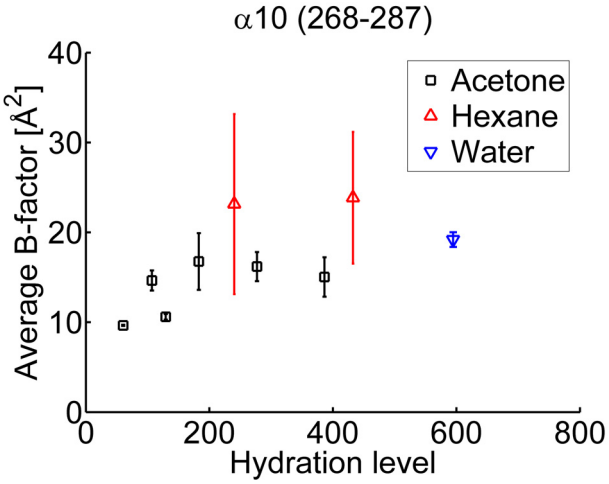


Figure 3.17: B-factor averaged over residues 268–287, as obtained from simulations in acetone (black), hexane (red triangles) and pure water (blue triangles). Standard error estimates were based on 3–5 replica simulations which were started from different initial velocities.

cules at the protein surface could be counteracting flexibility instead of promoting it. The fewer of these water molecules present, the more flexible the protein would be. The solvent would thus affect the flexibility indirectly by determining the number of slowly exchanged water molecules. In order to validate this hypothesis against the results presented here, the number of slowly exchanged water molecules of each simulation was estimated. A B-factor was calculated for each individual water molecule i in the simulation box using Equation (3.2). Here, the vector $\boldsymbol{r}_i$ refers to the position of the O atom of the water molecule. Water molecules with B-factors less than 25 Å^2 were classified as "slowly exchanged". The 25 Å^2 cutoff was used since this is a value comparable to the B-factors of non-hydrogen atoms of the more flexible sections of CALB. The numbers of slowly exchanged water molecules observed in the different systems are listed in Table 3.7. In Figure 3.18, the average B-factor of CALB is plotted vs. the number of slowly exchanged water molecules. The trend was followed fairly well, although the results obtained in the h286 simulations seemed to divert from the trend.

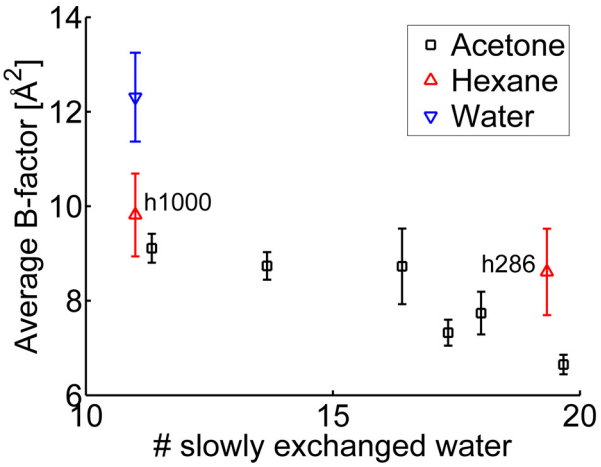


Figure 3.18: B-factor averaged over residues (11–307) vs. the number of slowly exchange water molecules, from the simulations carried out in acetone (black squares), hexane (red triangles) and pure water (blue triangles). Standard error estimates were based on values from 3-5 replica simulations started from different initial velocities (see Table 3.2). Note that the number of slowly exchanged water molecules generally decreases with decreasing hydration level. For hexane, it is marked which data point corresponds to which hydration level.

3.6 Summary

An MD study of CALB in water, one polar and one non-polar organic solvent at several hydration levels has been described. In the organic solvents, the structure and flexibility of CALB depended not only on the organic solvent as seen previously

(Trodler and Pleiss, 2008), but also on the hydration level. The medium had, in particular, profound effects on the structure and dynamics of two α-helices, $\alpha 5$ and $\alpha 10$, which are located on the rim of the active site pocket. The dynamics of these helices might be relevant for active site accessibility and substrate binding (Uppenberg *et al.*, 1994, 1995; Pleiss *et al.*, 1998; Skjøt *et al.*, 2009). The lid-like behavior reported by Skjøt *et al.* (2009) was however not seen here.

Structure and dynamics of CALB in organic solvent depends evidently on the hydration level. It is thus essential to control or measure this quantity, if the simulations carried out in different organic solvents are to yield properties that can be compared on a quantitative basis. In this chapter, hydration level has been quantified as the number of water molecules in the first solvation shell. This is straightforward, but not the only way to define hydration level, and it might be that other approaches are more appropriate. The water associated with the protein is for instance not necessarily contained in the first shell, and it might be that a more appropriate definition of hydration level should include several shells.

As stated in previous chapters, a rigorous approach to control hydration in experiments is to fix the thermodynamic water activity of the medium (Halling, 1989, 1990b; Valivety *et al.*, 1992b,a; Halling, 1994; Bell *et al.*, 1997). Implementing such an approach for MD simulations would significantly strengthen their applicability for studying enzymes in organic media, and would facilitate comparisons of protein properties obtained from simulations in different organic solvents. It would also make such simulation studies more compatible with experiments carried out at controlled water activity. A study of CALB structure and dynamics in organic solvents investigating the effects of water activity is described in Chapter 7.

4

Calculation of Water Activity

The previous chapters have already mentioned the importance of water activity as a parameter of non-aqueous biocatalytic system. This chapter discusses different approaches to consider this parameter in molecular dynamics (MD) simulations. Two main strategies to study how protein properties depend on water activity are here considered. The strategies are termed "Real-time" control and "*A posteriori*" analysis and are respectively described in Sections 4.1 and 4.2. The former comprises simulations of the protein in a non-aqueous medium in which the number of water molecules is automatically adjusted such that the desired water activity is maintained. In the latter strategy, conventional MD simulations are carried out like in Chapters 3 and 7, and the system water activity is calculated through post-analysis of the simulations. This approach will be employed for the protein simulations presented in Chapter 7 and is therefore more thoroughly described in this Chapter than the "Real-time" control approach is. In this work, the activity is evaluated using a methodology based on fluctuation solution theory (FST) (Kirkwood and Buff, 1951; O'Connell, 1971b,a), which is outlined in Sections 4.2.1–4.2.3.

The intention of these developments is not only to describe how thermodynamic activities are calculated, but also to take a step towards a general and efficient computational methodology that future protein simulation studies can be based on. The study of Branco *et al.* (2009) seems to be the only simulation study explicitly considering water activity as a parameter. In their study, the medium was however assumed to be an ideal gas mixture. This chapter will deal with the challenges arising when the medium is a non-ideal liquid mixture, as is the case for mixtures of water and organic solvents.

4.1 "Real-Time" Control of Water Activity?

In the MD simulations of *Candida antarctica* lipase B in acetone described in Chapter 3, the hydration water distributed itself between the protein surface and the bulk medium. This is likely to be a general behavior of simulations of proteins in water and organic solvent, except for simulations with very hydrophobic solvents (like hexane). By assessing the water content of the bulk medium it is, in principle, possible to evaluate the system water activity. One does however not know in advance how many water and organic solvent molecules to include in the simulation box in order to reach a specific water activity. One option is to guess the appropriate number of molecules and accept the resulting water activity. Another alternative is to tweak the water activity through a "trial-and-error" approach.

A more elegant approach would be to control the water activity "on-the-fly", similarly to how temperature and pressure are controlled in an NPT simulation. The

water activity a_{w}, is defined by how much the chemical potential of water in solution differs from the chemical potential in pure water

$$k_{\mathrm{B}}T \ln a_{\mathrm{w}} = \mu_{\mathrm{w}}(\boldsymbol{x}, T, P) - \mu_{\mathrm{w},0}(T, P) \tag{4.1}$$

where, $\mu_{\mathrm{w}}(\boldsymbol{x}, P, T)$ denotes the chemical potential of water in a solution of composition $\boldsymbol{x}$ at temperature T and pressure P. $\mu_{\mathrm{w},0}(P, T)$ denotes the chemical potential of pure water in a standard state which here is assumed to be pure, liquid water at temperature T and pressure P. Controlling the water activity is thus equivalent to controlling μ_{w}, which can be achieved in μVT-ensemble simulations using MD or Monte Carlo (MC). Such simulations involve insertion and removal of molecules, which is a standard procedure in MC (Allen and Tildesley, 1987). This is also possible in MD and is often termed grand canonical ensemble molecular dynamics (GMD) (Ji *et al.*, 1992; Lynch and Pettitt, 1997). The method requires treating one particle as "fractional" and employing special algorithms for selecting where to insert the new particles or which particles to remove. It seems however that no available MD software supports GMD. Another issue is that μ_{w} and P are to be fixed simultaneously, while fixing the number of non-aqueous molecules, which seems to be a non-standard application of GMD. Specifying the chemical potential corresponding to a desired activity requires furthermore knowledge of $\mu_{\mathrm{w},0}(T, P)$. One would have to evaluate this quantity from separate simulations of pure water.

Another possible approach is based on the Gibbs-ensemble Monte Carlo (GEMC) technique. In GEMC, molecules are distributed over multiple simulation boxes which usually realize different phases. Most commonly, the simulations comprise two boxes denoted I and II. In addition to the moves employed in regular MC simulations (e.g. single-molecule translation, rotation and conformational change, and simulation box shrinking/growing), moves are introduced that allow molecules to be transferred between the boxes (Panagiotopoulos, 1987b,a; Panagiotopoulos *et al.*, 1988). These moves equilibrate the two boxes with an external heat bath and piston of temperature T and P, respectively. The transfer moves furthermore ensure that the chemical potentials of each molecular species are equal in both boxes. Thus

$$\begin{aligned} T^I &= T^{II} = T \\ P^I &= P^{II} = P \\ \mu_i^I &= \mu_i^{II}, i = 1, \ldots, v \end{aligned} \tag{4.2}$$

where v denotes the number of molecular species. It is however possible to restrict the transfer moves to certain molecular species. In this case, only the chemical potential of these species will be equilibrated. For the other species, the number of molecules present in each box remains constant throughout the simulation.

Simulating a protein in an organic medium at a fixed a_{w} would according to Equation (4.1) require that $\mu_{\mathrm{w}}(\boldsymbol{x}, T, P)$ is fixed relative to $\mu_{\mathrm{w},0}(T, P)$. Naturally, the simulation should comprise two boxes, as shown in Figure 4.1. Box I would contain the protein, organic solvent molecules and hydration water. Box II would contain only water molecules and would thus have a water activity of unity. Transfer moves would be applied to water molecules exclusively. In order to fix a specific $a_{\mathrm{w}} < 1$ in box I, the acceptance probability for transfer moves would be modified. Panagiotopoulos *et al.* (1988) derived the acceptance probability for transferring a

molecule of type i from box I to box II in terms of insertion and removal MC moves in the grand canonical ensemble. An attempted removal of a molecule in box I is accepted with the probability

$$P_{\text{acc}} = \min\left[\frac{N_i^I}{V^I}\exp\left(-(\mu_i^I + \Delta U^I)/k_{\text{B}}T\right), 1\right] \tag{4.3}$$

where N_i^I, V^I and ΔU^I denote respectively the number of particles of type i in box I prior to the move, the volume of box I and the difference in configurational potential energy induced by the move. μ_i^I denotes the chemical potential of species i in box I which is a control parameter in the μVT ensemble. Likewise, an attempted insertion of a molecule in box II is accepted with the probability

$$P_{\text{acc}} = \min\left[\frac{V^{II}}{(N_i^{II}+1)}\exp\left((\mu_i^{II} - \Delta U^{II})/k_{\text{B}}T\right), 1\right] \tag{4.4}$$

Considering a molecule transfer move from box I to box II as composed of a removal and an insertion, the corresponding acceptance probability is given by (Panagiotopoulos *et al.*, 1988)

$$P_{\text{acc}} = \min\left[\frac{V^{II}N_i^I}{V^I(N_i^{II}+1)}\exp\left(-(\Delta U^I + \Delta U^{II} + \Delta\mu_i)/k_{\text{B}}T\right), 1\right] \tag{4.5}$$

where $\Delta\mu_i \equiv \mu_i^I - \mu_i^{II}$. In conventional GEMC simulation, $\Delta\mu_i$ is set to zero. It seems, in principle, possible to use a non-zero value, which will fix the difference in chemical potential in the two boxes. For the setup shown in Figure 4.1, setting $\Delta\mu_{\text{w}} = k_{\text{B}}T\ln a_{\text{w}}$ yields a water activity of a_{w} in box I at equilibrium, since the water activity of box II containing pure liquid water is unity.

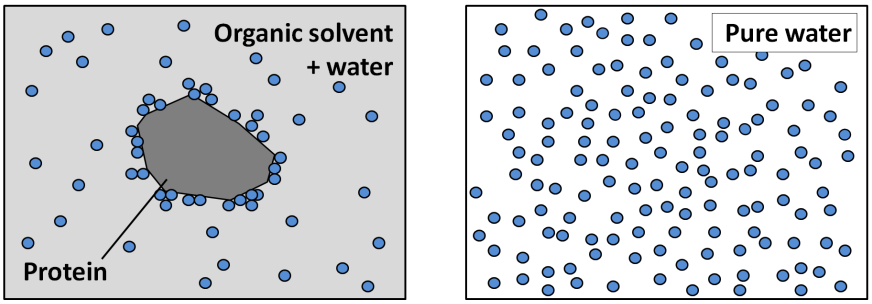


Figure 4.1: GEMC approach to controlling the water activity in a protein simulation. Box I contains protein, organic solvent and hydration water. Box II contains only water molecules. Only water molecules are transferred between the two boxes, and the transfer probability is weighted in order to fix a specific water activity in box I.

There is at least one GEMC package, MCCCS Towhee[1] (Martin and Siepmann, 1999), that allows to include proteins. Its capabilities are however somewhat limited

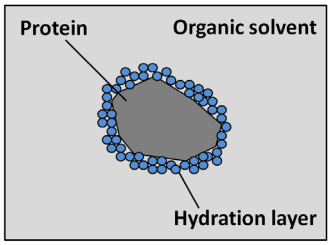
[1] http://towhee.sourceforge.net/

for this particular application. First, efficient MC sampling of protein conformations requires diligent selection of MC moves and assignment of appropriate weights, which in the case of protein simulation is a time-consuming procedure (see e.g. Hu *et al.* (2006)). Second, and more detrimental, is that MC packages usually lack support of for execution on parallel machines. Hence, current MC methods seem to be impractical for simulations of large systems such as those described in Chapter 3.

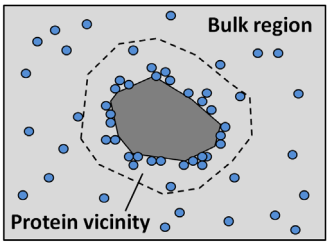
4.2 *A Posteriori* Analysis Approach

An alternative to "real-time" control of the water activity is to run the MD simulations with a fixed number of water and organic solvent molecules, as was done in Chapter 3, and evaluate the water activity *a posteriori*. This is less elegant since the number of water and organic solvent molecules that will result in the desired activity needs to be estimated or simply guessed, when the simulation is set up. Nevertheless, this approach seems to be more feasible than the "real-time" control approach discussed in the previous section.

With this "*a posteriori*" approach, the protein simulation needs to be sufficiently long, such that the partitioning of water molecules between the bulk medium and protein vicinity is equilibrated (see Figure 4.2). Considering the bulk phase as a binary mixture of water and organic solvent, the water molecule fraction in the bulk phase x_{w} is calculated, and the water activity is obtained as $a_{\mathrm{w}} = \gamma_{\mathrm{w}}(x_{\mathrm{w}})x_{\mathrm{w}}$, where $\gamma_{\mathrm{w}}(x_{\mathrm{w}})$ denotes the water activity coefficient at the composition x_{w}.



(a)



(b)

Figure 4.2: Simulation of a protein in a water-miscible organic solvent. (a) The water molecules are initially located near the protein surface (a). (b) After equilibration, some water molecules have mixed with the solvent. The bulk region refers to the region where the composition of water/organic solvent is homogeneous and the molecular distribution is (approximately) unaffected by the presence of the protein. Protein vicinity refers to the region around the protein where the solvent composition is different from the bulk composition, due to different preferential binding interactions of water and organic solvent molecules. In Chapter 7, a distance 10 Å from the protein surface defines the boundary between the protein vicinity and the bulk region.

This requires that $\gamma_{\mathrm{w}}(x_{\mathrm{w}})$ is evaluated. It is here of interest to study the significance of the water activity as driving force for protein hydration in simulation. It

is therefore not optimal to use correlations of experimental data, such as UNIFAC (Hansen *et al.*, 1991), in order to get $\gamma_w(x_w)$ since there might be discrepancies between simulation results and experimental data, which are due to the force field. More appropriate is to evaluate $\gamma_w(x_w)$ from separate simulations of the binary water/organic solvent mixture using the same force field as with the protein simulations.

GEMC (Panagiotopoulos, 1987b,a; Panagiotopoulos *et al.*, 1988) is probably the most applied molecular simulation method for computing thermodynamic properties of mixtures of relatively small molecules. In order to evaluate activity coefficients using GEMC, one would simulate the co-existing vapor and liquid phases of the binary mixture. Two boxes would be used containing respectively the vapor and liquid phases. Such simulations would be carried out at several compositions, ranging from pure water to pure organic solvent, which would yield the vapor pressure as a function of composition. A model for the excess Gibbs energy per molecule (G^E) would be fitted to the vapor pressure curve, which would allow the activity coefficients to be extracted (Smith *et al.*, 2005). This procedure assumes that the activity coefficients are independent of pressure, which usually is reasonable. A drawback using the GEMC method for this purpose is that one probably needs to carry out the time-consuming process of selecting and weighting MC moves specifically for each system one desires to simulate. GEMC furthermore requires that the vapor phase is simulated explicitly, which seems superfluous in this context, since it is only the properties of the liquid phase which are sought. A third limitation is that simulation of two phases in equilibrium requires extensive sampling. This is inevitably time consuming since MC is difficult to parallelize and no GEMC software that can be run on parallel machines seems to be available.

Christensen *et al.* (2007c,b,a) developed an alternative method for computing thermodynamic properties of mixtures which only requires that the liquid phase is simulated. The method is based on FST (Kirkwood and Buff, 1951; O'Connell, 1971b,a), which relates derivatives of thermodynamic functions to integrals of the pair radial distribution functions (RDFs). The binary mixture is simulated at a few compositions, and the pair RDFs are calculated. The RDFs are integrated to yield thermodynamic derivative properties to which a G^E model is fitted. The method is applicable when only properties of the liquid phase are needed (as is the case here), or when the vapor phase can be modeled by simpler means, e.g. as an ideal gas or by a virial expansion. Since the approach is based entirely on post-analysis of simulation trajectories it can be applied with MD as well as with MC and allows thus the user to exploit efficient (i.e. parallelized) MD software for running the simulations. The method, which is described below, is employed in this work to analyze MD results for water/organic solvent mixtures in order to calculate the activity coefficients.

4.2.1 Fluctuation Solution Theory

FST is a framework of equations linking the microscopic structure of a fluid with its macroscopic thermodynamic properties (Kirkwood and Buff, 1951; O'Connell, 1971b,a). Although FST is valid for solutions with an arbitrary number of components, the following discussion is restricted to binary mixtures. The microscopic

structure is represented by the pair RDFs in the grand canonical ensemble, averaged over molecular orientations and conformations, $g_{ij}^{(\mu VT)}(r)$ (Figure 4.3). $g_{ij}^{(\mu VT)}(r)$ denotes the RDF for the molecular pair ij, which for a binary mixture with species 1 and 2 is be 11, 12 or 22. For convenience, the total correlation functions (TCFs), defined by $h_{ij}(r) = g_{ij}^{(\mu VT)}(r) - 1$, are introduced. The total correlation function integrals (TCFIs), denoted H_{ij}, are spatial integrals of the TCFs, defined by

$$H_{ij} = \rho \int_0^\infty r^2 h_{ij}(r) dr \tag{4.6}$$

where ρ is the molecular density of the system. The quantities H_{ij}, which sometimes are referred to as Kirkwood-Buff (KB) integrals, can conveniently be collected in a matrix

$$\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \tag{4.7}$$

Note that $H_{12} = H_{21}$. The TCFIs are related to thermodynamic properties through the exact equation (Kirkwood and Buff, 1951)

$$\mathbf{A} = (\mathbf{X} + \mathbf{XHX})^{-1} \tag{4.8}$$

where $\mathbf{X}$ is a diagonal matrix whose elements are given by mixture component molecule fractions, $X_{ii} = x_i$, and where the elements of $\mathbf{A}$ are given by

$$A_{ij} = \frac{N}{k_\mathrm{B}T} \left(\frac{\partial \mu_i}{\partial N_j} \right)_{T,V,N_k,k\neq j} \tag{4.9}$$

where N, N_j and V denote total particle number, particle number of component j and total system volume, respectively. Via thermodynamic transformation of Equation (4.9), the following equations can be derived (Kirkwood and Buff, 1951; O'Connell, 1971b)

$$\rho k_\mathrm{B} T \kappa_T = \frac{1 + x_1 H_{11} + x_2 H_{22} + x_1 x_2 \left(H_{11} H_{22} - H_{12}^2 \right)}{1 + x_1 x_2 \Delta H} \tag{4.10}$$

$$\rho \bar{v}_1 = \frac{1 + x_2 \left(H_{22} - H_{12} \right)}{1 + x_1 x_2 \Delta H} \tag{4.11}$$

$$\left(\frac{\partial \ln \gamma_1}{\partial x_1} \right)_{T,P,N_2} = \frac{-x_2 \Delta H}{1 + x_1 x_2 \Delta H} \tag{4.12}$$

where $\Delta H \equiv H_{11} + H_{22} - 2H_{12}$ and κ_T, $\bar{v}_1$ and γ_1 denote isothermal compressibility, partial molecular volume and activity coefficient for component 1, respectively.

4.2.2 Correlation Function Integrals from Simulation

The RDFs for the three molecular pairs of a binary mixture are straightforwardly obtained from MD simulation and could in principle be integrated numerically to yield the thermodynamic derivative properties of Equations (4.10)–(4.12). This task has proved to be more difficult expected. In order to evaluate the integral

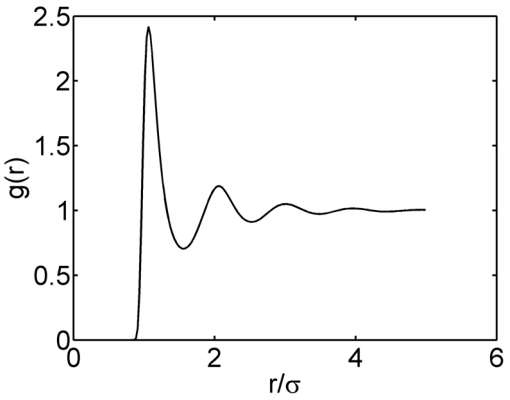


Figure 4.3: RDFs obtained from simulation of the pure Lennard-Jones (LJ) fluid at a temperature of 1.5 ϵ/k_{B} and a number density of 0.8 σ^{-3}, where ϵ and σ are the LJ potential depth and diameter, respectively. The simulations are described in Section 6.1. The function $g(r)$ is defined as the average density of molecular centers in a spherical shell at distance r from a reference molecule, relative to the overall density.

numerically, it is in practice necessary to impose an upper limit R_{lim} to the integral in Equation (4.6), which becomes

$$H_{ij}(R_{\mathrm{lim}}) = \rho \int_0^{R_{\mathrm{lim}}} r^2 h_{ij}(r) dr \tag{4.13}$$

R_{lim} needs to be chosen sufficiently large such that the integral converges within this distance. This means that $H_{ij}(R_{\mathrm{lim}})$ should be insensitive to further increases of R_{lim}. Experience shows however that $H_{ij}(R_{\mathrm{lim}})$ rarely converges within the range accessible in MD simulation. This is illustrated in Figure 4.4 and has been attributed to several factors. Firstly, the RDFs can only be obtained for values of r up to half the simulation box dimension. Obviously, the integral will not converge if the simulated system is too small. Increasing the system size might remedy this but will increase the required computational effort, which is undesirable. Secondly, the obtained RDFs might be inaccurate due to finite-size effects (Salacuse *et al.*, 1996) which could cause the divergence of the integrals. The finite-size effects include e.g. artifacts arising from the use of periodic boundary conditions. The finite-size effect that probably has the most detrimental impact arises however from the fact that the MD simulations are carried out in the isothermal-isobaric (NPT) ensemble, while the FST equations are derived for the grand-canonical (μVT) ensemble. Differences between properties obtained in these two ensembles are usually negligible, but this might not be the case for TCFIs, which can be understood from the original derivation of Equation (4.9). Independently of the ensemble, one has that (Kirkwood and Buff, 1951)

$$\rho \int_0^{\infty} r^2 (g_{ij}(r) - 1) dr = \frac{\langle N_i N_j \rangle - \langle N_i \rangle \langle N_j \rangle}{\langle N_i \rangle \langle N_j \rangle} - \delta_{ij} \tag{4.14}$$

where δ_{ij} denotes the Kronecker delta and $\langle\cdot\rangle$ denotes ensemble average. In the μVT ensemble, the number of particles of each species is a stochastic variable. The covariance expression on the right-hand side can then be related to the chemical potential derivatives of the $\langle N_i\rangle$:s, from which Equation (4.9) follows (Kirkwood and Buff, 1951). In the NPT ensemble, the particle numbers are fixed. The covariance term in Equation (4.14) is thus zero, and one obtains $H_{ij} = -\delta_{ij}$. This suggests that $g_{ij}^{(\mu VT)}(r)$ and $g_{ij}^{(NPT)}(r)$ are sufficiently dissimilar to yield significantly different integrals. It has however been argued that the difference mainly is manifested in that $g_{ij}^{(NPT)}(r) \nrightarrow 1$ for large r (Ben-Naim, 2008; Gray and Gubbins, 1984). The two functions should thus coincide well except for when r is large, as argued by Salacuse *et al.* (1996).

The difficulties arising when attempting to evaluate the integral in Equation (4.6) have been recognized by many researchers, and several approaches to overcome these difficulties have been proposed. A few of these approaches are discussed below.

Truncation Approach The simplest strategy to is the truncation approach by Weerasinghe and Smith (2003), in which the integrals are evaluated using a specific upper limit R_{lim}. Since the TCFIs do not converge within the range sampled in simulation, the results are sensitive to the choice of R_{lim}. It is therefore advised to average $H(R_{\text{lim}})$ with R_{lim} varying in a selected interval. There seems to be no guidelines for how to select this interval, other than that it should correspond to one oscillation of the TCFs.

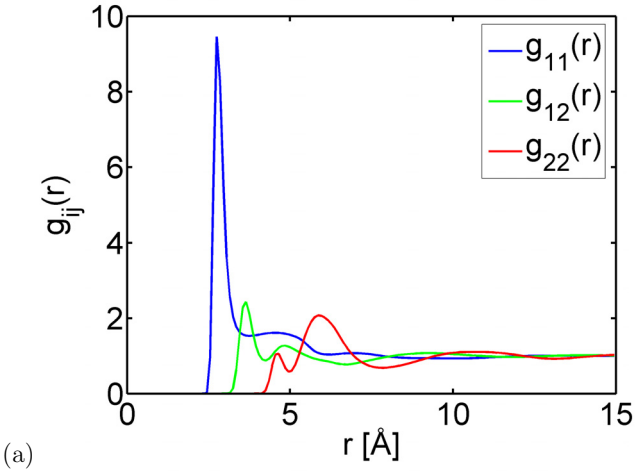
RDF Shifting Approaches The approaches by Perera and Sokolić (2004) and Hess and van der Vegt (2009) both attempt to correct the RDFs obtained from simulation by rescaling them according to

$$g_{ij}^{*}(r) = \alpha_{ij} g_{ij}(r) \tag{4.15}$$

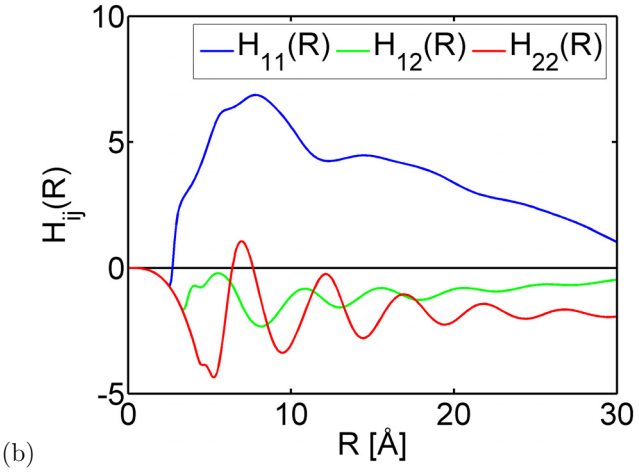
α_{ij} is chosen in order to enforce that $g_{ij}^{*}(r)$ approaches unity at long distances. In the Perera approach, this is done by "brute force", namely by requiring that $g_{ij}^{*}(R_{\text{lim}}) = 1$ at a specific R_{lim} within the sampling range. Thus, $\alpha_{ij} = (g_{ij}(R_{\text{lim}}))^{-1}$ (Perera and Sokolić, 2004). The Hess approach is based on the realization that due to that the number of molecules is fixed, the fluid composition far from a given molecule is different from the overall composition. In order to correct for this, one employs a scaling factor given by

$$\alpha_{ij} = \frac{V_{\text{box}} - V(R_{\text{lim}})}{V_{\text{box}} - V(R_{\text{lim}}) - \int_0^{R_{\text{lim}}} (g_{ij}(r) - 1) r^2 dr - \frac{\delta_{ij}}{x_i \rho}} \tag{4.16}$$

where V_{box}, $V(R_{\text{lim}})$, x_i, ρ and δ_{ij} denote respectively the simulation box average volume, the volume of a sphere of radius R_{lim}, molecule fraction of species i, molecular density and the Kronecker delta, respectively. The scaling factor is evaluated using a distance R_{lim} beyond which $g_{ij}(r)$ should be roughly constant. There does not seem to be any systematic way to choose the parameter R_{lim} either for the Perera and Sokolić (2004) or Hess and van der Vegt (2009) approach. Although these two methods are straightforward to implement, they do apparently not eliminate the necessity of selecting an appropriate truncation distance. This is a limitation, since the results are sensitive to this selection.



(a)



(b)

Figure 4.4: (a) RDFs obtained from simulation of mixture of water (1) and t-butanol (2) at $x_1 = 0.65$, temperature of 323 K and pressure of 1 atm. The simulations are described in Section 6.4. (b) Numerically evaluated integrals $H_{ij}(R_{\text{lim}})$ (see Equation (4.13)) are shown. These integrals do not converge within the sampled range. The line at $H_{ij}(R_{\text{lim}}) = 0$ is included to guide the eye.

Tail-Modeling Approaches Christensen *et al.* (2007a) explored the possibility to obtain convergent integrals by replacing the long-range part of $g_{ij}(r)$ with an empirical model. The authors initially employed the model of Matteoli and Mansoori (1995), but abandoned this for a simpler model better suited for the task, given by

$$g_{ij}^{\text{model}}(r) = a \cdot \exp\left(-b(r-c)\right) \sin\left(d(r-c)\right) \tag{4.17}$$

where a, b, c and d are adjustable parameters. In order to evaluate the TCFI, the integral is split into "direct" and "indirect" contributions

$$H_{ij} = \int_0^{r_{u3}} r^2 \left(g_{ij}(r) - 1\right) dr + \int_{r_{u3}}^{\infty} r^2 \left(g_{ij}(r) - 1\right) dr \tag{4.18}$$

where r_{u3} denotes the third unity of $g_{ij}(r)$. The first, "direct" term is evaluated numerically. In the second, "indirect" term, $g_{ij}(r)$ is replaced by the model expression (Equation (4.17)). The four parameters are determined by fitting the expression to the indirect part of the RDF obtained from simulation and the indirect contribution to H_{ij} is evaluated by analytically integrating the model expression.

The success of the method was demonstrated through a number of case studies (Christensen *et al.*, 2007c,b). The method was in a later study applied to evaluate isothermal compressibilities of pure alkane fluids (Wedberg *et al.*, 2008). The procedure as described here did not yield satisfactory results, but a modified approach was however found to perform better. In this approach, a tail model corresponding to the anti-derivative of Equation (4.17) was fitted to the tail of the function $H(R_{\text{lim}})$, i.e. the truncated numerical integral of $g(r)$ as a function of the upper integration limit. The tail model was then used to extrapolate $H(R_{\text{lim}})$ to $R_{\text{lim}} = \infty$ which yielded the value of the TCFI. This approach has however only been successfully tested on pure fluids.

Either in its original formulation or in the modification by Wedberg *et al.* (2010), the approach by Christensen *et al.* (2007a) is only applicable for systems where the TCF tails can be approximated by Equation (4.17). This is not the case in general. Examples of systems where the TCF shows a different behavior are mixtures where one component is water. Such mixtures are of central importance in this work.

Fourier Space Approach Finally, an interesting approach was proposed by Nichols *et al.* (2009). It does not rely on correcting $g_{ij}(r)$ but to directly evaluate the structure factors $S_{ij}(k)$, which are related to the RDFs via the radial Fourier transform (Allen and Tildesley, 1987)

$$S_{ij}(k) = 1 + 4\pi\rho \int_0^{\infty} r^2 \frac{\sin kr}{kr} \left(g_{ij}(r) - 1\right) dr \tag{4.19}$$

The TCFIs are in turn related to the structure factors via $H_{ij} = S_{ij}(0) - 1$. This allows the target derivative properties to be expressed "as functions of k", whose values for $k = 0$ are the actual values of the properties. The inaccuracies in the RDFs at large r are reflected in $S_{ij}(k)$ at small k. In particular, $S_{ij}(0)$ cannot be reliably obtained directly from simulation. The solution is to extrapolate the k-dependent property functions to $k = 0$ by fitting them to polynomials, whose degrees are selected by empirical means.

Good results were obtained for Lennard-Jones mixtures (Nichols *et al.*, 2009), but the method has seemingly not yet been tested for molecular fluids. The user is furthermore still required to select fitting polynomials, as well as the interval in k-space where the fitting is to be done.

Summary In conclusion, various methods for computing TCFIs from molecular simulations have been proposed. All these methods involve selecting a truncation

radius or employing an empirical model. It seems that a robust and general approach that can be directly applied to simulations of a new mixture is yet to be developed. In Chapters 5–6, an attempt is made to develop a robust and theoretically well-motivated method for correcting and extending the TCFs obtained from simulation, such that accurate integrals can be obtained.

4.2.3 Regression of Molecular Gibbs Energy Models

With the FST approach by Christensen *et al.* (2007c,b,a), simulations of the binary mixture are carried out at various compositions. The TCFIs obtained from the simulations are via Equation (4.12) converted into activity coefficient derivatives. The modified Margules (mM) model for the Gibbs energy per molecule is defined by (Abbott and van Ness, 1975)

$$\frac{G^{\mathrm{E}}}{k_{\mathrm{B}}Tx_1x_2} = A_{21}x_1 + A_{12}x_2 - \frac{\alpha_{12}\alpha_{21}x_1x_2}{\alpha_{12}x_1 + \alpha_{21}x_2 + \eta x_1x_2} \tag{4.20}$$

where A_{21}, A_{21}, α_{12}, α_{21} and η are adjustable parameters. The activity coefficient derivatives are obtained by differentiating this expression twice, according to (Smith *et al.*, 2005)

$$\left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P,N_2} = N\left(\frac{\partial \ln \gamma_1}{\partial N_1}\right)_{T,P,N_2} = \frac{N}{k_{\mathrm{B}}T}\left(\frac{\partial^2 (NG^{\mathrm{E}})}{\partial N_1^2}\right)_{T,P,N_2} \tag{4.21}$$

The model in Equation (4.20) is fitted to the derivatives obtained from simulation by minimization of the objective function (Christensen *et al.*, 2007c)

$$SS = \sum_i \frac{\left[\left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{\mathrm{MD},i} - \left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{\mathrm{mM},i}\right]^2}{\sigma_i^2} \tag{4.22}$$

where subscript i denotes the mixture composition, subscripts MD and mM denote that the quantity is obtained from MD simulation and the model, respectively, and σ_i denotes the standard error in the activity coefficient derivative estimated by the MD simulation of composition i.

The precise use of the mM model is typically adapted to the complexity of the mixture. Either a two-parameter version (α_{12}, α_{21} and η set to zero), four-parameter version (η set to zero) or five-parameter version (all parameters allowed to be non-zero) is employed. In this work, all three versions of the model are tested for fitting the MD data. The four-parameter version is preferred over the two-parameter version if it fits the MD data significantly better as indicated by the smallness of the minimized objective function. Likewise, the five-parameter version is preferred over the four-parameter version if it results in a significantly better fit. If switching to a version of higher complexity does not result in a significant improvement, the simpler version is preferred.

A more sophisticated approach to selecting the appropriate version considers the standard errors in the parameters obtained from the fitting procedure (Abbott and van Ness, 1975). The simple approach described above is however adequate for the present application.

Once the appropriate model version has been selected and the parameters are determined, the activity coefficients are obtained from the model by differentiation of Equation (4.20) (Smith *et al.*, 2005)

$$\ln \gamma_1 = \frac{1}{k_\mathrm{B} T} \left(\frac{\partial (N G^\mathrm{E})}{\partial N_1} \right)_{T,P,N_2} \tag{4.23}$$

4.3 Summary

Different approaches for calculating the water activity in non-aqueous protein simulations have been discussed in this chapter. "Real-time" control is the more elegant approach but seems to be too inefficient to be practical with current state-of-the-art simulation techniques. *A posteriori* analysis seems more feasible but requires efficient methods to evaluate the activity coefficients of binary liquid mixtures. FST analysis of MD simulations of mixtures is a promising candidate for this. This method is however still somewhat limited due to the present lack of robust methods to obtain the TCFIs from simulation.

5

Modeling the Direct Correlation Function

The total correlation functions (TCFs) of a liquid usually show damped oscillating behavior at long distances. Their integrals are the differences between the total positive and negative areas, which often are large and of similar magnitude. Modeling the long-range behavior of the TCF is therefore expected to be difficult, since the integral is likely to be sensitive to errors in the tail. It was suggested by Rowlinson (1965) and later by O'Connell (1971b) that the direct correlation function (DCF), which will be defined in Section 5.2, is more likely to be accurately predicted since it is of shorter range than the TCF, as illustrated in Figures 5.1(a)–(b). The fluctuation solution theory (FST) relations for thermodynamic derivatives given in the previous chapter can be reformulated in terms of integrals of the DCFs (O'Connell, 1971b,a), and several successful corresponding state theories for derivative properties rely on modeling these integrals, in place of the total correlation function integrals (TCFIs) (Brelvi and O'Connell, 1972; Huang and O'Connell, 1987; Abildskov *et al.*, 2009, 2010a).

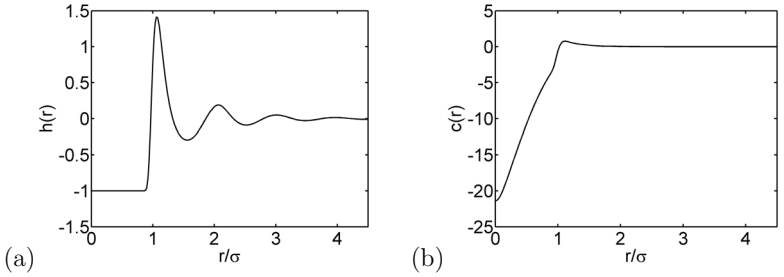


Figure 5.1: (a) TCF and (b) DCF obtained from simulation of the pure Lennard-Jones (LJ) fluid at a temperature of 1.5 ϵ/k_{B} and a number density of 0.8 σ^{-3}, where ϵ and σ are the LJ potential depth and diameter, respectively. The simulations are described in Section 6.1. The DCF has a simpler behavior than the TCF when r is large.

This chapter explores the possibility of utilizing approximations for the DCFs to improve molecular dynamics (MD) simulation estimates of the TCFIs. The computational methodology is based on a method due to Verlet (1968), which extends the TCFs obtained from simulation by enforcing that the corresponding DCFs at large spatial separation follows an approximate expression. Since it seems that the

method in previous studies only has been applied to the pure Lennard-Jones (LJ) fluid (Verlet, 1968) and LJ mixtures (Jolly *et al.*, 1976), the developments of this chapter are focused on how the method can be extended to molecular fluid mixtures.

Sections 5.1–5.2 describes elementary properties of the TCFs, DCFs and the Ornstein-Zernike (OZ) equation that relates these two correlation functions. The computational methodology is described in Section 5.3, and an efficient and relatively straightforward numerical implementation is proposed. In Section 5.4, approximate expressions for the behavior of the DCFs at long range are derived from statistical mechanical fluid theory. Although these approximations are based on well-known results, there seems to be no previous study that discusses the long-range DCF of molecular fluids at the same level of detail.

The discussion of this chapter is restricted to the theoretical aspects and implementation details of the method. In Chapter 6, numerical tests of the method are carried out for the sake of validation.

5.1 Molecular Correlation Functions

The TCFs, $h_{ij}(r)$, and the pair radial distribution functions (RDFs), $g_{ij}(r)$, were in Chapter 4 introduced as functions of the spatial distance r between the centers of mass (COM) of two molecules (hence the term *radial* distribution function for $g_{ij}(r)$). In general, molecular correlation functions are as well functions of the molecular orientations, ω_1 and ω_2.

In order to represent orientations, molecule 1 is associated with an intrinsic coordinate system (x_1, y_1, z_1), which is fixed with respect to the geometry of the molecule. Molecule 2 is similarly associated with a coordinate system (x_2, y_2, z_2). A reference coordinate system (X, Y, Z) is also introduced, which either could be fixed in space, or defined such that the Z-axis is parallel to the vector separating the COM of molecules 1 and 2. These two representations are respectively termed the *space-fixed* and *intermolecular frame* representations. The orientation ω_1 of molecule 1 is represented by the three-dimensional rotation mapping the system (X, Y, Z) to (x_1, y_1, z_1) which e.g. can be represented in terms of Euler angles $\omega_1 \equiv (\phi_1\theta_1\chi_1)$ (Gray and Gubbins, 1984). The orientation ω_2 of molecule 2 is represented likewise.

The orientation-dependent TCF pairing molecules of species i and j is written as $h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)$, using the space-fixed representation, where $\boldsymbol{r}_{12}$ is the vector separating the COM of molecules 1 and 2. If the intermolecular frame representation is used, this vector is by definition parallel with the Z-axis, and can thus be represented by just its magnitude r_{12}, in which case the TCF is written as $h_{ij}(r_{12}\omega_1\omega_2)$.

It is for the following discussion convenient to split the TCFs into isotropic and anisotropic parts.

$$h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = h_{ij}(r_{12}) + h_{ij}^{(\mathrm{a})}(\boldsymbol{r}_{12}\omega_1\omega_2) \tag{5.1}$$

The isotropic part $h_{ij}(r_{12})$ is identical to the radial TCF of Chapter 4, and is obtained from the orientation-dependent TCF by averaging out the angular dependence

$$h_{ij}(r_{12}) \equiv \langle h_{ij}(r_{12}\omega_1\omega_2)\rangle_{\omega_1\omega_2} \tag{5.2}$$

where

$$\langle\cdot\rangle_{\omega_1} \equiv \frac{1}{8\pi^2}\int d\omega_1 \equiv \frac{1}{8\pi^2}\int_0^{2\pi} d\phi_1 \int_{-1}^{1} d(\cos\theta_1) \int_0^{2\pi} d\chi_1 \tag{5.3}$$

The anisotropic part $h_{ij}^{(\mathrm{a})}(\boldsymbol{r}_{12}\omega_1\omega_2)$ is defined such that Equation (5.1) is satisfied and fulfills therefore

$$\left\langle h_{ij}^{(\mathrm{a})}(\boldsymbol{r}_{12}\omega_1\omega_2)\right\rangle_{\omega_1\omega_2} = 0 \tag{5.4}$$

If flexible molecules are considered, the TCFs are formally also functions of the conformations of the molecules 1 and 2. Since the organic molecules considered in this work are rather small and do not possess multiple conformations (with exception for n-hexane), this dependency is suppressed throughout this work.

5.2 The Ornstein-Zernike Equation

In a theoretical study of critical opalescence, Ornstein and Zernike (1914) introduced the DCF which in contrast to the TCF remained of short range, and its integral remained finite as the critical point was approached. For this reason, Ornstein and Zernike (1914) fount that the formulae for the scattering intensity were more conveniently expressed in terms of the DCF than in terms of the TCF. The DCF for the component pair ij in a molecular fluid mixture is here denoted by $c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)$, and is in general a function of molecular orientations, as well as the COM separation vector. The DCFs are related to the TCFs via the Ornstein-Zernike (OZ) equation (Gray and Gubbins, 1984)

$$\begin{aligned} h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) &= c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) + \\ &\rho\sum_l x_l \int \langle h_{il}(\boldsymbol{r}_{13}\omega_1\omega_3) c_{lj}(\boldsymbol{r}_{32}\omega_3\omega_2)\rangle_{\omega_3}\, d\boldsymbol{r}_3 \end{aligned} \tag{5.5}$$

where ρ and x_l denote respectively the number density of the fluid and the molecule fraction of component l. Equation (5.5) can be viewed as the definition of the DCFs. Alternatively, they can be defined via a cluster series expansion in terms of the Mayer f-function (McQuarrie, 1976), or via the functional derivative of an external potential acting on the fluid molecules, with respect to the local fluid density (Lebowitz and Percus, 1963b).

In analogy with Equations (5.1), (5.2), and (5.4), the DCF can be written as a sum of isotropic and anisotropic parts

$$c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = c_{ij}(r_{12}) + c_{ij}^{(\mathrm{a})}(\boldsymbol{r}_{12}\omega_1\omega_2) \tag{5.6}$$

Substituting Equations (5.1) and (5.6) into Equation (5.5) and averaging out the angular dependence leads to

$$\begin{aligned} h_{ij}(r) &= c_{ij}(r) + \rho\sum_l x_l \int h_{il}(r_{13}) c_{lj}(r_{32}) d\boldsymbol{r}_3 \\ &+\rho\sum_l \int \left\langle \left\langle h_{il}^{(\mathrm{a})}(\boldsymbol{r}_{13}\omega_1\omega_3)\right\rangle_{\omega_1} \left\langle c_{lj}^{(\mathrm{a})}(\boldsymbol{r}_{32}\omega_3\omega_2)\right\rangle_{\omega_2}\right\rangle_{\omega_3} d\boldsymbol{r}_3 \end{aligned} \tag{5.7}$$

If the last term in Equation (5.7) is neglected, a simplified version of the OZ equation is obtained, in which the isotropic DCFs and TCFs are related independently of the anisotropic terms

$$h_{ij}(r) = c_{ij}(r) + \rho\sum_l x_l \int h_{il}(|\boldsymbol{r}-\boldsymbol{r'}|) c_{lj}(r') d\boldsymbol{r'} \tag{5.8}$$

The computational approach to compute the long range behavior of the TCFs described and tested in this chapter is based on Equation (5.8) rather than (5.5), since Equation (5.8) is far more convenient to implement.

How can the neglection of the anisotropic term in Equation (5.7) be justified? There seems to be no rigorous arguments for this that holds in the general case. There is however several results that make the assumption seem plausible. Firstly, Equation (5.8) is exact in several integral equation theories of fluids with anisotropic interactions, such as the mean-spherical approximation and the generalized mean-field theory (Gray and Gubbins, 1984). Secondly, Wang *et al.* (1973) showed by Monte Carlo simulations that anisotropic forces have a rather small impact on the isotropic TCF. This applied to systems with interaction potentials whose isotropic part was of the LJ type and whose anisotropic part was either of the dipole-dipole or quadrupole-quadrupole type, even when the corresponding multipole moments were large. The same result was argued by Gubbins and O'Connell (1974) using a perturbation theory approach. Thirdly, several corresponding-states theories for thermodynamic derivative properties which are based on modeling the DCFIs (Gubbins and O'Connell, 1974; Brelvi and O'Connell, 1972; Huang and O'Connell, 1987) employ a macroscopic version of the approximation of Equation (5.8), described in detail by O'Connell (1994). The approximation has not been found to limit the success and applicability of the theories.

Anisotropic interactions influence the isotropic TCF as well as the isotropic DCF. It should be stressed that the treatment based on Equation (5.8) does not neglect effects such as these, but rather constitutes a simplified way of relating the two isotropic correlation functions.

It should finally be stressed that the approximation of Equation (5.8) can be systematically improved by considering the spherical harmonic expansions of the orientation-dependent TCFs and DCFs (Gray and Gubbins, 1984), for which the isotropic correlation functions constitute the first terms. By representing the orientation-dependent correlation functions by several spherical harmonic terms, one obtains a more accurate approximation of the full OZ equation. The significance of this improvement is however not known at this stage and is likely to depend on the system studied. Investigating the improvements obtained by considering several spherical harmonics would be very beneficial, but is due to time limitations beyond the scope of this work. The study is nevertheless recommended to be part of future investigations.

5.3 The Verlet Method

The DCFs can in principle be calculated from the TCFs obtained from molecular simulations. These TCFs are however only obtained for a finite spatial range and might be inaccurate at large separations due to issues discussed in Section 4.2.2. This seems to cause the calculation of the DCFs to be unstable since the errors in the TCFs at large separations are amplified and propagated to the entire spatial range. This seems to be the case especially at high densities. The computational method described in this section offers a more stable approach to calculate the DCFs from the TCFs.

5.3.1 Method Formulation

The method due to Verlet (1968) aims at correcting the TCFs obtained from simulation for possible finite-size effects, such as those summarized by Salacuse *et al.* (1996), and to extend them to long range. The method was originally introduced in a study of the pure LJ fluid, focusing of the qualitative behavior of the DCF and the structure factor, and was later applied to an LJ mixture (Jolly *et al.*, 1976). With the present formulation of the method, one seeks to numerically determine TCFs and DCFs that satisfy the OZ equation (5.8), under the constraints

$$\begin{cases} h_{ij}(r) &= h_{\mathrm{MD},ij}(r), & r \leq R_{ij} \\ c_{ij}(r) &= t_{ij}(r), & r > R_{ij} \end{cases} \tag{5.9}$$

where $h_{\mathrm{MD},ij}(r)$ are the TCFs obtained from simulation, $t_{ij}(r)$ are approximations of the long range part of the DCFs and R_{ij} are the distances where the TCFs from simulation are matched with the extensions to be calculated. Solving the OZ equation under these constraints yields TCFs extended to arbitrary separations. These TCFs can be integrated numerically to yield the TCFIs. Alternatively, the DCFs which also are obtained through the procedure can be integrated to yield the DCFIs, from which the corresponding TCFIs can be computed. The procedure of solving the OZ equation with the given constraints requires explicit approximations $t_{ij}(r)$ for the behavior of the DCFs at large separations. Such approximations are derived in Section 5.4. One furthermore needs to select appropriate matching distances R_{ij}, which is elaborated in Section 5.5.

5.3.2 Implementation

Commonly, the Wiener-Hopf factorization technique is applied when the DCF is computed numerically from the TCF or vice versa (Gray and Gubbins, 1984; Jolly *et al.*, 1976; Ramirez *et al.*, 2005). For the present application, employing the Fourier-transformed OZ equation turned out to be a numerically feasible approach, and relatively straightforward to implement. The following discussion is restricted to fluid mixtures that have at most two components.

Applying the three-dimensional Fourier transform to Equation (5.8) transforms the convolution into a product

$$\tilde{h}_{ij}(k) = \tilde{c}_{ij}(k) + \rho \sum_{l=1}^{2} \tilde{h}_{il}(k) x_l \tilde{c}_{lj}(k) \tag{5.10}$$

where $\tilde{h}_{ij}(k)$ denotes the Fourier transformation of $h_{ij}(r)$, which due to the radial symmetry is reduced to the zero:th-order Hankel transform, defined by (Gray and Gubbins, 1984)

$$\tilde{h}_{ij}(k) = 4\pi \int_0^\infty dr r^2 \frac{\sin(kr)}{kr} h_{ij}(r) \tag{5.11}$$

and likewise for $\tilde{c}_{ij}(k)$. The function $h_{ij}(r)$ is recovered from the inverse Hankel transform and is given by

$$h_{ij}(r) = \frac{4\pi}{(2\pi)^3} \int_0^\infty dk k^2 \frac{\sin(kr)}{kr} \tilde{h}_{ij}(k) \tag{5.12}$$

Utilizing that $\tilde{h}_{12}(k) = \tilde{h}_{21}(k)$ and $\tilde{c}_{12}(k) = \tilde{c}_{21}(k)$, Equation (5.10) can be written as a linear system

$$\tilde{\mathbf{h}}(k) = (\mathbf{I} + \rho\underline{\mathbf{H}}(k))\,\tilde{\mathbf{c}}(k) \tag{5.13}$$

with

$$\tilde{\mathbf{h}}(k) = \begin{pmatrix} \tilde{h}_{11}(k) \\ \tilde{h}_{12}(k) \\ \tilde{h}_{22}(k) \end{pmatrix} \quad \tilde{\mathbf{c}}(k) = \begin{pmatrix} \tilde{c}_{11}(k) \\ \tilde{c}_{12}(k) \\ \tilde{c}_{22}(k) \end{pmatrix} \tag{5.14}$$

and

$$\underline{\mathbf{H}}(k) = \begin{pmatrix} x_1\tilde{h}_{11}(k) & x_2\tilde{h}_{12}(k) & 0 \\ 0 & x_1\tilde{h}_{11}(k) & x_2\tilde{h}_{12}(k) \\ 0 & x_1\tilde{h}_{12}(k) & x_2\tilde{h}_{22}(k) \end{pmatrix} \tag{5.15}$$

Equations (5.11)–(5.13) provide a route for computing $c_{ij}(r)$ given $h_{ij}(r)$. The function $h_{ij}(r)$ is Hankel-transformed to yield $\tilde{h}_{ij}(k)$. The linear system in Equation (5.13) is then solved for $\tilde{c}_{ij}(k)$ for each k, and the inverse Hankel transform is applied to obtain $c_{ij}(r)$. Solution of the problem of Equation (5.9) requires that the long-range part of $h_{ij}(r)$ is adjusted until the long-range part of $c_{ij}(r)$ concurs with the theoretical result $t_{ij}(r)$. This is here accomplished by a Newton iteration scheme for which grids in r and k space are introduced

$$r_\alpha \equiv \alpha \cdot \Delta r, \alpha = 0, \ldots, N_r \tag{5.16}$$

$$k_\beta \equiv \beta \cdot \Delta k, \beta = 0, \ldots, N_k \tag{5.17}$$

which also implies the upper cutoffs $R_\mathrm{c} = N_r \cdot \Delta r$ and $K_\mathrm{c} = N_k \cdot \Delta k$ for the integrals in Equations (5.11) and (5.12), respectively. Note that R_c is *not* the sampling limit set by the simulation box dimensions, but should typically be set much larger than this. The TCFs, DCFs and their Hankel transforms are at the current iteration step t represented discretely by vectors $\mathbf{h}_{ij}^{(t)}$, $\mathbf{c}_{ij}^{(t)}$, $\tilde{\mathbf{h}}_{ij}^{(t)}$ and $\tilde{\mathbf{c}}_{ij}^{(t)}$, with elements defined by

$$h_{ij,\alpha}^{(t)} \equiv h_{ij}^{(t)}(r_\alpha), \alpha = 1, \ldots, N_r \tag{5.18}$$

$$c_{ij,\alpha}^{(t)} \equiv c_{ij}^{(t)}(r_\alpha), \alpha = 1, \ldots, N_r \tag{5.19}$$

$$\tilde{h}_{ij,\beta}^{(t)} \equiv \tilde{h}_{ij}^{(t)}(k_\beta), \beta = 1, \ldots, N_k \tag{5.20}$$

$$\tilde{c}_{ij,\beta}^{(t)} \equiv \tilde{c}_{ij}^{(t)}(k_\beta), \beta = 1, \ldots, N_k \tag{5.21}$$

Equation (5.11) for the TCF is approximated by truncating the integral at R_c and using the trapezoidal rule

$$\tilde{\mathbf{h}}_{ij}^{(t)} = \mathbf{T}\mathbf{h}_{ij}^{(t)} \tag{5.22}$$

where the elements of the matrix $\mathbf{T}$ are given by

$$T_{\beta\alpha} = 4\pi\Delta r r_\alpha^2 \frac{\sin k_\beta r_\alpha}{k_\beta r_\alpha}\left(1 - \frac{\delta_{0\alpha} + \delta_{N_r\alpha}}{2}\right) \tag{5.23}$$

with $\alpha = 1, \ldots, N_r$, $\beta = 1, \ldots, N_k$, $\delta_{\alpha\alpha'}$ denoting the Kronecker delta, and where it is understood that $\sin k_\beta r_\alpha / k_\beta r_\alpha$ becomes unity if either k_β or r_α vanishes. Equation (5.12) for the DCF is approximated in a similar way by truncating the integral at K_c

$$\mathbf{c}_{ij}^{(t)} = \mathbf{U}\tilde{\mathbf{c}}_{ij}^{(t)} \tag{5.24}$$

with

$$U_{\alpha\beta} = \frac{4\pi}{(2\pi)^3}\Delta k k_\alpha^2 \frac{\sin k_\beta r_\alpha}{k_\beta r_\alpha}\left(1 - \frac{\delta_{0\beta} + \delta_{N_k\beta}}{2}\right) \tag{5.25}$$

with α and β like in Equation (5.23). As stated above, the middle step converting $\tilde{\mathbf{h}}_{ij}^{(t)}$ to $\tilde{\mathbf{c}}_{ij}^{(t)}$ is carried out by solving the linear system of Equation (5.13) for each value of β.

Now, let n_{ij} denote the index such that $r_{n_{ij}} \leq R_{ij} < r_{n_{ij}+1}$, and let $\underline{\mathbf{h}}_{ij}^{(t)}$ and $\underline{\mathbf{c}}_{ij}^{(t)}$ denote vectors containing the elements of $\mathbf{h}_{ij}^{(t)}$ and $\mathbf{c}_{ij}^{(t)}$, respectively, with $n_{ij}+1 \leq \alpha \leq N_r$. At each iteration step, $\underline{\mathbf{h}}_{ij}^{(t)}$ is updated according to

$$\underline{\mathbf{h}}_{ij}^{(t+1)} = \underline{\mathbf{h}}_{ij}^{(t)} + \Delta\underline{\mathbf{h}}_{ij}^{(t)} \tag{5.26}$$

where $\Delta\underline{\mathbf{h}}_{ij}^{(t)}$ according to the Newton method is found by solution of the linear system

$$\begin{pmatrix} \mathbf{J}_{11}^{11} & \mathbf{J}_{12}^{11} & \mathbf{J}_{22}^{11} \\ \mathbf{J}_{11}^{12} & \mathbf{J}_{12}^{12} & \mathbf{J}_{22}^{12} \\ \mathbf{J}_{11}^{22} & \mathbf{J}_{12}^{22} & \mathbf{J}_{22}^{22} \end{pmatrix} \begin{pmatrix} \Delta\underline{\mathbf{h}}_{11}^{(t)} \\ \Delta\underline{\mathbf{h}}_{12}^{(t)} \\ \Delta\underline{\mathbf{h}}_{22}^{(t)} \end{pmatrix} = \begin{pmatrix} \Delta\underline{\mathbf{c}}_{11}^{(t)} \\ \Delta\underline{\mathbf{c}}_{12}^{(t)} \\ \Delta\underline{\mathbf{c}}_{22}^{(t)} \end{pmatrix} \tag{5.27}$$

where the right-hand side represents the difference between the approximation of the long-range DCF to be enforced and the currently computed DCF

$$\Delta\underline{c}_{ij,\alpha}^{(t)} \equiv t_{ij}(r_\alpha) - \underline{c}_{ij,\alpha}^{(t)}, \qquad \alpha = n_{ij}+1, \ldots, N_r \tag{5.28}$$

$\mathbf{J}_{i'j'}^{ij}$ denotes the Jacobian for the transformation mapping $\Delta\underline{\mathbf{h}}_{i'j'}^{(t)}$ to $\Delta\underline{\mathbf{c}}_{ij}^{(t)}$, i.e.

$$\mathbf{J}_{i'j'}^{ij} \equiv \begin{pmatrix} \partial c_{ij,n_{ij}+1}^{(t)}/\partial h_{i'j',n_{i'j'}+1}^{(t)} & \cdots & \partial c_{ij,n_{ij}+1}^{(t)}/\partial h_{i'j',N_r}^{(t)} \\ \vdots & \ddots & \vdots \\ \partial c_{ij,N_r}^{(t)}/\partial h_{i'j',n_{i'j'}+1}^{(t)} & \cdots & \partial c_{ij,N_r}^{(t)}/\partial h_{i'j',N_r}^{(t)} \end{pmatrix} \tag{5.29}$$

The calculation of these Jacobians is described in Appendix C. Note that the short range part of the discretized TCFs remains constant throughout the iteration. The short range part of the calculated DCFs is not used within the iteration scheme. The short range part of the DCF obtained from the final iteration is however considered for the selection of the parameters R_{ij}, as will be explained in Section 5.5.

A MATLAB implementation of the Newton scheme was used for all calculations in this work. Initially, the discretized TCFs were set to $h_{ij,\alpha}^{(0)} = h_{\mathrm{MD},ij}(r_\alpha)$ for all r_α within the sampling range for $h_{\mathrm{MD},ij}(r)$, and $h_{ij,\alpha}^{(0)} = 0$ for larger α. The iteration was carried out until the criterion

$$\sum_{i,j}\sum_{\alpha=n_{ij}+1}^{N_r}\left(\Delta c_{ij,\alpha}^{(t)} r_\alpha^2\right)^2 < \eta \tag{5.30}$$

with η set to 10^{-4} or less, was fulfilled. Typically, this was achieved after 5-15 iterations. For some of the simulated systems, in particular those at high density where the functions $h_{ij}(r)$ had significant structure beyond the sampling range, this initial guess was too poor and the iteration consequently diverged. The initial guess

could however be improved by a heuristic approach. Instead of setting $h^{(0)}_{ij,\alpha}$ to zero beyond the sampling range, it was equated with the tail model by Christensen *et al.* (2007a), i.e. Equation (4.17) of Section 4.2.2. The model parameters were determined by fitting the model to the sampled $h_{\mathrm{MD},ij}(r)$, as described in Section 4.2.2. Using this approach, the Newton iteration converged for all studied systems.

5.4 Approximating the Long-Range DCF

The benefit of considering the DCFs in place of the TCFs is that the former exhibit simpler behavior at long distances, and is expected to decay monotonically when the COM separation distance is larger than a few molecular diameters. In this section, an approximate expression for the DCF at large separations is derived. The idea is to express the angle averaged DCF as a truncated series expansion in negative powers of the separation distance r. In fact, only the slowest decaying term proportional to r^{-6} will be retained.

A well-known result for the long-range part of the DCF states that (Gray and Gubbins, 1984)

$$c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \to -\beta u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2), r_{12} \to \infty \tag{5.31}$$

where $u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)$ is the pair interaction potential for molecules of type i and j and $\beta \equiv (k_{\mathrm{B}}T)^{-1}$. For the present discussion, it is assumed that the potential is similar to the intermolecular part of the CHARMM force field (MacKerell Jr. *et al.*, 1998), which is given in Appendix A. Such a potential is the sum of the LJ and Coulombic interactions between the individual atoms. Since uncharged molecules are considered here, the Coulombic part is at large separation dominated by the dipole-dipole interaction. The potential is written as

$$u_{ij} = u^{(\mathrm{LJ})}_{ij} + u^{(\mathrm{dd})}_{ij} \tag{5.32}$$

At large separations, the LJ term $u^{(\mathrm{LJ})}_{ij}$ is $O(r_{12}^{-6})$. The dipole-dipole term $u^{(\mathrm{dd})}_{ij}$ is $O(r_{12}^{-3})$, but vanishes if one averages out the orientational dependence.

Equation (5.31) is derived from the cluster series expansion for the DCF, of which the first two terms are (McQuarrie, 1976)

$$\begin{aligned} c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = {} & f_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) + \\ & \rho f_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \sum_l \int x_l \left\langle f_{il}(\boldsymbol{r}_{13}\omega_1\omega_3) f_{lj}(\boldsymbol{r}_{32}\omega_3\omega_2) \right\rangle_{\omega_3} d\boldsymbol{r}_3 \\ & + \ldots \end{aligned} \tag{5.33}$$

where $f_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \equiv \exp\left(-\beta u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)\right) - 1$ defines the Mayer f-function. This function, which constitutes the first term on the right-hand side of Equation (5.33), approaches the potential times $-\beta$ at large separations, which can be realized by Taylor-expanding the exponential. The second term of the cluster expansion is a function that decays as $O(f_{ij}^2)$. This is due to that the decay of a convolution is of the same order as that of the slowest decaying factor, and that the convolution here once again is multiplied with the Mayer f-function. Lebowitz and Percus (1963a) argued that all higher order terms of the cluster expansion for a similar reason decay

as $O(f_{ij}^2)$. Thus, the DCF is for large separations dominated by the first term, i.e. the Mayer f-function, which decays according to Equation (5.31). This also implies that the long-range behavior of the DCF is insensitive to the density.

It should however be noted that Equation (5.31) is valid for the orientational-dependent DCF. One might expect that the angle-averaged DCF is related to the angle-averaged potential in a similar way, i.e.

$$c_{ij}(r_{12}) \to -\beta u_{ij}(r_{12}), r_{12} \to \infty \tag{5.34}$$

This is however not necessarily true in general, in particular not for potentials like Equation (5.32). The reason is that the anisotropic part dominated by the dipole-dipole term decays slower than the isotropic part. While the dipole-dipole term vanishes when the angular dependence is integrated out, it may give rise to a second-order contribution to the DCF which does not vanish. Such a contribution decays as $O(r_{12}^{-6})$, i.e. the same order as the LJ interaction.

Consider for instance the DCF of a fluid with molecular interactions given by Equation (5.32). In the low-density limit, the DCF approaches the Mayer f-function. If the exponential is Taylor-expanded, one obtains

$$\begin{aligned} f_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = &-\beta u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) - \beta u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2) \\ &+\tfrac{\beta^2}{2}\left(u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2)\right)^2 + O(r_{12}^{-9}) \end{aligned} \tag{5.35}$$

The second-order contribution from the dipole-dipole term needs to be retained since it is $O(r_{12}^{-6})$. If the orientational dependence is integrated out, the first-order contribution of the dipole-dipole interaction vanishes, while the second-order contribution becomes

$$\left\langle \left(u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2)\right)^2 \right\rangle_{\omega_1\omega_2} = -\frac{2\mu_i^2\mu_j^2}{3r_{12}^6} \tag{5.36}$$

where μ_i denotes the magnitude of the dipole moment of a molecule of type i. This expression is derived in Appendix D. It is in fact identical to the Keesome potential, which is an effective spherically symmetric potential which approximates the dipole-dipole interaction at low density (Reed and Gubbins, 1973). Due to this contribution, it is clear that the asymptotic behavior of Equation (5.34) is not satisfied for this type of molecular interactions.

At high density, calculations might be even more complicated, since the higher-order terms of the cluster expansion (Equation (5.33) are significant. These terms might contain second-order contributions from the dipole-dipole interaction, which affect the DCF decay. It seems unlikely that one will be able to derive exact expressions for the leading order of the DCF decay at high density. An approximation of the decay can however be obtained using the hypernetted chain (HNC) relation. The HNC relation is an approximation of the DCF in terms of the TCF and the intermolecular potential, given by (McQuarrie, 1976)

$$\begin{aligned} c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = &-\beta u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) + h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \\ &- \log\left(1 + h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)\right) \end{aligned} \tag{5.37}$$

For a pure fluid composed of rigid, dipolar molecules, the asymptotic behavior of the TCF was by Nienhuis and Deutch (1971) shown to follow

$$h(\boldsymbol{r}_{12}\omega_1\omega_2) \to -\frac{\beta G^2}{\epsilon} u^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2), r_{12} \to \infty \tag{5.38}$$

where ϵ and G respectively denote dielectric constant and Kirkwood factor of the fluid. The latter is defined by

$$G \equiv \frac{\langle \boldsymbol{\mu} \cdot \boldsymbol{M} \rangle}{\mu^2} \tag{5.39}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{M}$ respectively denote the dipole moment of a single molecule in the fluid and the total dipole moment of the fluid, and where $\langle \cdot \rangle$ denotes ensemble average. The Kirkwood factor also appears in the formula for the dielectric constant, which with "tinfoil" boundary conditions is given by $\epsilon = 1 + 4/3\pi\beta\rho\mu^2 G$ (Allen and Tildesley, 1987).

The result of Nienhuis and Deutch (1971) (Equation (5.38)) applies to pure fluids. It is however straightforwardly extended to mixtures, for which it takes the form

$$h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \to -\frac{\beta G_i G_j}{\epsilon} u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2), r_{12} \to \infty \tag{5.40}$$

where G_i is a specific Kirkwood factor, defined by

$$G_i \equiv \frac{\langle \boldsymbol{\mu}_i \cdot \boldsymbol{M} \rangle}{\mu_i^2} \tag{5.41}$$

where $\boldsymbol{\mu}_i$ denotes the dipole moment of a single molecule of type i.

Combining this asymptotic result with the HNC relation (Equation (5.37)) and Taylor-expanding the logarithm of the latter, leads to

$$\begin{aligned} c_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) &= -\beta u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) + h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) \\ &\quad - \log\left(1 + h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)\right) \\ &= -\beta u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) + \frac{1}{2}\left(h_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2)\right)^2 + O(r_{12}^{-9}) \\ &= -\beta u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) - \beta u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2) \\ &\quad - \frac{\beta^2 G_i^2 G_j^2}{2\epsilon^2}\left(u_{ij}^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2)\right)^2 + O(r_{12}^{-9}) \end{aligned} \tag{5.42}$$

It was in the last step utilized that the pair potential is the sum of LJ and dipole-dipole contributions. The asymptotic behavior of the isotropic DCF is obtained by integrating out the angular dependence. This is carried out explicitly in Appendix D for systems where the LJ term is given as in CHARMM (MacKerell Jr. *et al.*, 1998). To leading order, the result is

$$c_{ij}(r_{12}) = t_{ij}(r_{12}) + O(r_{12}^{-8}) \tag{5.43}$$

with

$$t_{ij}(r_{12}) \equiv -2\beta\left[\sum_{\alpha\in M_i, \beta\in M_j} \epsilon_{\alpha\beta} R_{\min,\alpha\beta}^6\right] r_{12}^{-6} + \frac{\beta^2 G_i^2 G_j^2 \mu_i^2 \mu_j^2}{3\epsilon^2} r_{12}^{-6} \tag{5.44}$$

where M_i denotes the set of atoms of a molecule of type i and $\epsilon_{\alpha\beta}$ and $R_{\min\alpha\beta}$ denote CHARMM parameters for the LJ interaction between atoms of type α and β.

Equation (5.44) is in this work applied to approximate the DCF tail as required by the Verlet method of Section 5.3. The coefficient of the first (LJ) term of Equation (5.44) is evaluated directly from the CHARMM parameters. The coefficient of the second (dipole-dipole) term requires that G_i, G_j and ϵ are evaluated from the simulations, which is a straightforward application of Equation (5.41), and using (Allen and Tildesley, 1987)

$$\epsilon = 1 + \frac{4\pi\beta\rho}{3}\sum_i x_i\mu_i^2 G_i \tag{5.45}$$

which however requires that the simulations are carried out with electrostatic forces evaluated using tinfoil boundary conditions.

The derivation of Equation (5.44) required that the HNC relation (Equation (5.37)) was employed. One could of course as well have employed another closure relation such as the Percus-Yevick relation. This would have lead to another result, namely one with the coefficient of the dipole-dipole term of Equation (5.44) given by

$$\frac{(2G_iG_j - 1)\,\beta^2\mu_i^2\mu_j^2}{3\epsilon} \tag{5.46}$$

The HNC result is nevertheless preferred for several reasons. Firstly, this is because integral equation theories for dipolar fluids based on the HNC relation are more well explored in the literature than those based on the Percus-Yevick relation, and their accuracy is more well documented (Murad *et al.*, 1983; Fries and Patey, 1985; Rossky, 1985). A second reason is that in the light of Equations (5.31) and (5.35), it seems reasonable to interpret the DCF at large separations as the negative of an effective pair potential (times β). One would thus expect the dipole-dipole contribution to be positive. The HNC result (Equation (5.44)) is positive definite while the Percus-Yevick result (Equation (5.46)) is not. In fact, the Percus-Yevick derived coefficient becomes negative for several of the fluid mixtures studied in Chapter 6. The HNC result thus appear more reliable, and Equation (5.44) is used to approximate the DCF tail within this work. [1].

5.5 Determining Matching Distance

Figures 5.2(a)–(b) compare the DCF $c(r)$ evaluated directly from the TCF obtained from simulation of the pure LJ fluid, truncated at half the box dimension, with $c(r)$ obtained via the Verlet method using two different matching distances R. The appearance of Figure 5.2(a) is typical for a DCF calculated from a truncated TCF, which often deviates from the theoretical r^{-6} decay. Similar observations have been reported in previous studies (Ramirez *et al.*, 2005) and probably arise from the same deficiencies in TCFs that cause divergence of the TCFIs. With the Verlet method, the TCFs are corrected at long distances such that the r^{-6} decay in $c(r)$ is

[1] The journal paper [Wedberg, O'Connell, Peters and Abildskov, *Mol. Simul.*, 2010] was submitted and accepted before the developments of this section were made, and thus apply less rigorous approximations for the DCF tail. The simulations described in the paper have however been re-analyzed using the approximations of this section for their presentation in Chapter 6.

enforced. With $R = 4\sigma$, $c(r)$ does not decay monotonically, but becomes negative before the positive tail approximation is recovered at 4σ (Figure 5.2(b), solid line). With $R = 2.5\sigma$, $c(r)$ is more smooth and decays essentially monotonically, with a slight oscillation (Figure 5.2(b), dashed line). The behavior of $c(r)$ at small r, and the DCF peak seem to be insensitive to the choice of R. In simulations of mixtures, similar trends are observed for all three DCFs.

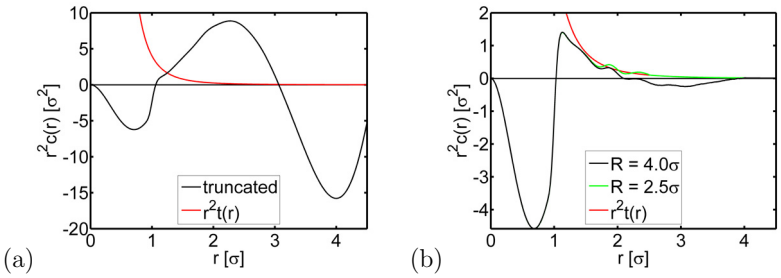


Figure 5.2: DCF evaluated from TCF obtained from simulation of the LJ fluid at $\rho = 0.85$ (see Section 6.1). In (a), the DCF is calculated directly from the truncated TCF (black line), while in (b), it is evaluated from the TCF extended by the Verlet method using $R = 4\sigma$ (green line) or $R = 2.5\sigma$ (black line). The tail model $t(r)$ is indicated in both plot (red line).

A robust approach for selecting the parameters R_{ij} is crucial for obtaining accurate properties from the resulting correlation functions. The original approach by Verlet (1968) who studied the pure LJ fluid was to choose R as one of the zeros of $h(r)$ (i.e. R such that $h(R) = 0$), which was thought to ensure continuity of the extended correlation functions. It was also reported that the integral of the extended TCF was insensitive to which zero R was set to, obviously except for the first zero. The analysis of the pure LJ and Stockmayer fluids presented in Section 6.1 in this work did however not entirely support these claims, since the TCFI generally depended on R. The integrals were nevertheless approximately constant when R was chosen between 1.5 and 2.2 σ, where σ was the LJ diameter. This corresponded well to the location of the third zero of $h(r)$, which however is unlikely to be a universal result that extends to molecular fluids. Selecting R at one of the zeros of $h(r)$ does not seem to ensure continuity of the extended TCF. Continuity of the TCF appears instead to depend on whether the DCF is matched continuously with the tail approximation $t(r)$.

The parameter R_{ij} is here chosen by a procedure that attempts to match the DCFs continuously with the tail approximation. The calculations are first carried out using preliminary parameter values, $R^{\dagger}_{ij}$, set, for instance, to the largest radius for which $h_{ij}(r)$ is sampled. Preliminary DCFs $c^{\dagger}_{ij}(r)$ are calculated using the Verlet method. The parameters R_{ij} for the final calculations are then chosen at a point after the peak of $c^{\dagger}_{ij}(r)$ where $c^{\dagger}_{ij}(r)$ intersects the tail approximation $t_{ij}(r)$. For the DCF shown in Figure 5.2, this occurs at $r = 1.88\sigma$. In case $t_{ij}(r)$ does not intersect

$c_{ij}^{\dagger}(r)$, R_{ij} is chosen as the value of r after the peak which minimizes the expression

$$\frac{\left|c_{ij}^{\dagger}(r) - t_{ij}(r)\right|}{|t_{ij}(r)|} \tag{5.47}$$

Since the peak of $c_{ij}(r)^{\dagger}$ is approximately independent of the matching distances, this procedure yields values of R_{ij} that are insensitive to the preliminary parameters $R_{ij}^{\dagger}$, as long as these are large enough for $c_{ij}^{\dagger}(r)$ to be past its peak. For all systems studied in Chapter 6, this procedure was employed for selecting R_{ij}. It was confirmed by visual inspection that the extended TCFs and DCFs obtained were continuous.

5.6 Summary

A computational method for correcting TCFs obtained from molecular simulation for finite-size effects and extending them to large separations has been described. The method relies on enforcing that the corresponding DCFs follow theoretical approximations at long distances. The OZ equation which relates the TCFs with the DCFs was simplified by assuming that the isotropic and anisotropic parts of the correlation functions decouple. Approximations for the long-range behavior have been derived from statistical mechanical theory of fluids, and a relatively straightforward numerical implementation was presented.

This chapter has focused on the theoretical aspects of the method while the next chapter is devoted to numerical tests for the sake of validation.

6

Testing the Verlet Method

This chapter describes a set of molecular dynamics (MD) simulations carried out in order to test the capabilities of the Verlet method described in the previous chapter. The ultimate goal of this methodology is to enable efficient simulation-based prediction of thermodynamic properties that are in good agreement with experimental data. This however comprises two tasks. First, a reliable integration procedure must be established. As discussed in Section 4.2.2, this always involves assumptions or approximations of the total correlation functions (TCFs). It is thus crucial to validate integration results against previous simulations or alternative simulation-based routes to the same properties. Second, potential models and parameters for the relevant atoms and chemical groups need to be optimized such that experimental properties are reproduced by simulation. Focus is here on the first step. The second step, i.e. force field development, is beyond the scope of this work.

The test systems fall into four categories as shown in Table 6.1 and those will be discussed in separate sections. For each test system, the methodology of Chapter 5 is applied to evaluate the total correlation function integrals (TCFIs). Depending on the system the TCFIs are converted into different thermodynamic derivative properties. For instance, for the simulations of pure fluids described in Sections 6.1 and 6.3, the isothermal compressibility is evaluated via Equation (4.10). These calculations are primarily verified by comparing the derivative properties obtained from the integration procedure with the same properties obtained from alternative analysis methods, or from simulation results in the literature. For the simulations of water/organic solvent mixtures described in Section 6.4, the derivative properties obtained by integration are also compared against values derived from correlations of experimental data. The consistency of these results depends thus not only on the accuracy of the integration procedure but also on the accuracy of the force field.

As discussed in Section 4.2.2, several methods for improving the accuracy of TCFIs have previously been proposed. Some of these methods are especially simple to implement namely the "truncation" method of Weerasinghe and Smith (2003), the "Hess" method of Hess and van der Vegt (2009) and the method of Perera and Sokolić (2004). The truncation and Hess methods are here applied to some of the test systems, and their performance is compared with the Verlet method. The method of Perera and Sokolić (2004) is omitted since it is very similar to the Hess method.

The simulations of water/organic solvent mixtures described in Section 6.4 are included here not only for verification of the integration methodology but also to develop excess Gibbs energy models for these mixtures that will be employed with the protein simulations of Chapter 7 in order to determine the bulk water activity.

Table 6.1: The four test cases described in this chapter. Simulations of pure and/or atomic fluids are considered in order to validate the integration method against previous simulations or alternative computational routes. The aqueous organic mixtures are employed as protein solvents in Chapter 7.

	Atomic fluids	Molecular fluids
Pure fluids	Lennard-Jones Stockmayer (Section 6.1)	n-alkanes 2-propanol water (Section 6.3)
Mixtures	Lennard-Jones/Stockmayer (Section 6.2)	water/acetone water/methanol water/t-butanol (Section 6.4)

6.1 Pure Lennard-Jones and Stockmayer Fluids

In this section, atomic fluids, i.e. fluids consisting of particles that are single atoms, are considered. The atoms interact via a pair potential which is the sum of Lennard-Jones (LJ) and electrostatic dipole-dipole interactions

$$u(\boldsymbol{r}_{12}\omega_1\omega_2) = u^{(\mathrm{LJ})}(r_{12}) + u^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2) \tag{6.1}$$

with

$$u^{(\mathrm{LJ})}(r_{12}) = 4\epsilon \left[\left(\frac{\sigma}{r_{12}} \right)^{12} - \left(\frac{\sigma}{r_{12}} \right)^{6} \right] \tag{6.2}$$

$$u^{(\mathrm{dd})}(\boldsymbol{r}_{12}\omega_1\omega_2)) = \frac{\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2}{r_{12}^3} - \frac{3(\boldsymbol{\mu}_1 \cdot \boldsymbol{r}_{12})(\boldsymbol{\mu}_2 \cdot \boldsymbol{r}_{12})}{r_{12}^5} \tag{6.3}$$

where $\boldsymbol{r}_{12}$, ω_1 and $\boldsymbol{\mu}_1$ denote respectively the separation vector of atoms 1 and 2, the orientation of atom 1 and the electric dipole moment vector of atom 1 which is a function of its orientation ω_1. In the LJ term, ϵ and σ denote respectively the LJ potential well depth and diameter. In the LJ fluid, all atoms have zero electric dipole moment and interact thus only via LJ forces. In the pure Stockmayer fluid, all atoms carry a non-zero dipole moment of magnitude μ.

The methodology of Chapter 5 was tested on these two fluids for several reasons. Firstly, these fluids are simple, and simulations data could therefore be acquired with relatively small computational effort. This allows for exploration of a rather wide range of system temperatures and densities. Secondly, the thermodynamics of these fluids obtained by simulations is well-documented in the literature. Thus, the thermodynamic derivative properties obtained from the extended pair-distribution function can relatively easy be validated against data derived from correlations of previous simulations (see Section 6.1.2). Thirdly, the simple form of the inter-atomic potentials in Equations (6.1)–(6.3) allows for a very basic test of the assumption that the Ornstein-Zernike (OZ) equation can be separated into isotropic and anisotropic parts as described in Section 5.2. The assumption is exact for the LJ fluid in which

anisotropic interactions are absent. For the Stockmayer fluid, the approximation is inexact, and it is expected to become less accurate as the dipole moment μ^2 increases since this parameter determines the strength of the anisotropic interactions. The accuracy of the properties obtained for large μ^2 might therefore indicate whether the simplified OZ equation (Equation (5.8)) is valid.

In this section and in the following (6.2), physical quantities are for convenience quoted in reduced units, where the quantities have been reduced with respect to the LJ parameters ϵ (energy), σ (length) and the atomic mass. Reduced quantities are marked with an asterisk (*).

6.1.1 Simulation Details

The simulations were carried out in the NVT ensemble (particle number, volume and temperature were constant) using 864 molecules initially arranged according to a face-centered cubic lattice and with initial velocities assigned according to the Maxwell-Boltzmann distribution of the corresponding temperature. The LJ parameters, σ and ϵ/k_B, were set to 3.405 Å and 119.8 K, respectively, and the molar mass, m, was set to 39.941 g/mol, which is the set of parameters commonly used to model argon. Periodic boundary conditions were employed in the x, y and z directions, and LJ forces were truncated at 4σ. The velocities were rescaled at each time step to maintain constant temperature.

In the LJ simulations, the Velocity Verlet algorithm (Allen and Tildesley, 1987) with a time step of 2 fs was employed to integrate the equations of motion. The systems were equilibrated for 100 ps and the production periods were 900 ps.

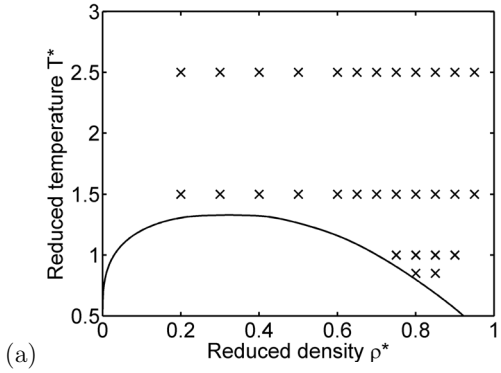
The simulations of the Stockmayer fluid were carried out using either $\mu^{*2} = 1$ or $\mu^{*2} = 3$, where $\mu^{*2} = \mu^2\epsilon^{-1}\sigma^{-3}$ denotes the reduced squared dipole moment. Electrostatic energies and forces were evaluated using the Ewald summation method with the parameter $\alpha = 3.5/L$ and the upper cutoff in reciprocal space, $n_\mathrm{c} = 6$, following the notation of Rapaport (2004). The fifth order Nordsieck-Gear predictor-corrector method (Allen and Tildesley, 1987) with a time step of 2 fs was employed to integrate the equations of motion. Rotational motion was implemented via the quaternion algorithm (Allen and Tildesley, 1987) for which all particles were assigned a reduced moment of inertia of 0.1. The dielectric constant was evaluated via the fluctuation in the total dipole vector (Allen and Tildesley, 1987). The systems were equilibrated for 100 ps, and the production periods were 900 ps.

A simulation program implemented in FORTRAN77 was used for the simulations. The temperatures and densities used in the simulations are shown in Figures 6.1(a)–(c).

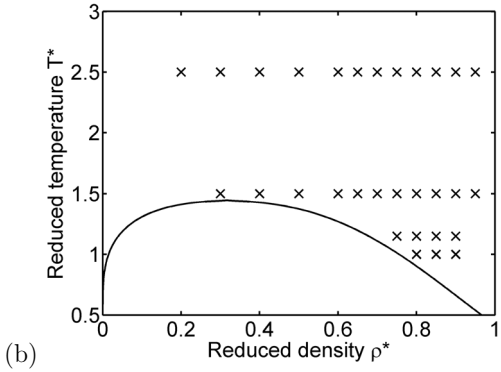
6.1.2 Results

For pure fluids, one has only one TCF $h(r)$ and one DCF $c(r)$. The isothermal compressibility κ_T which is a pure component property is expressed in terms of these functions as (cf. Equation (4.10))

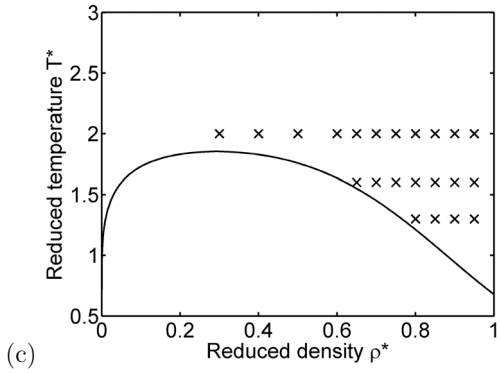
$$\rho k_\mathrm{B} T \kappa_T = 1 + 4\pi\rho \int r^2 h(r) dr \equiv 1 + H \tag{6.4}$$



(a)



(b)



(c)

Figure 6.1: State points used in the simulations (×) of (a) the LJ fluid, (b) the Stockmayer fluid with $\mu^{*2} = 1$ and (c) the Stockmayer fluid with $\mu^{*2} = 3$. The boundary of the vapor-liquid coexistence region is indicated for each system and was obtained from the equations of state of (a) Mecke *et al.* (1996)and (b, c) Gross and Vrabec (2006).

$$(\rho k_{\mathrm{B}} T \kappa_T)^{-1} = 1 - 4\pi\rho \int r^2 c(r) dr \equiv 1 - C \tag{6.5}$$

where ρ, k_{B} and T respectively denote density, Boltzmann constant and temperature, and H and C define respectively the TCFI and the direct correlation function integral (DCFI) of the pure fluid. Studying partial molecular volumes and activity coefficient derivatives which are mixture properties is irrelevant for pure fluids.

In order to assess how accurately the isothermal compressibility was obtained by the Verlet method it was also evaluated via the equations of state (EOS) by Mecke *et al.* (1996) and Gross and Vrabec (2006). The Mecke EOS was derived from a Helmholtz energy expression developed by correlating energy, temperature, pressure and density for a large number of MD and Monte Carlo (MC) simulations of the LJ fluid. The Gross and Vrabec EOS was developed in a similar way for a fluid with 2-center LJ and dipole-dipole interactions. The Stockmayer fluid is the special case of this fluid when the distance between the two LJ centers is zero. For both EOS, the isothermal compressibility was obtained by analytical differentiation. These EOS were employed rather than EOS fitted to experimental data due to reasons hinted above. Good agreement with experimental data requires not only that the Verlet method is valid, but also that the force field used in the simulations is accurate. Since the present goal is to validate the Verlet method rather than the accuracy of the force field, the choice of simulation-based EOS is more appropriate than experimental-based ones.

Overall, the pressures obtained from the present simulations of the LJ fluid were found to be in very good agreement with the Mecke *et al.* (1996) EOS. Deviations in pressure were typically around 0.2% and generally less than 1%. The Stockmayer simulations carried out with $\mu^{*2} = 1$ yielded pressures within 2% of those obtained from the Gross and Vrabec (2006) EOS. For $\mu^{*2} = 3$, typical discrepancies in pressure were around 2-3%, although some were as large as 5-6% for a few state points at low reduced temperature.

The radial distribution function (RDF) of a representative simulation at high density is shown in Figure 6.2(a). In Figure 6.2(b), the truncated integral of the same non-extended RDF is shown as function of upper integration limit. The integral does not converge within the sampling limit and as the sampling limit is approached, the oscillations in the integral are still large enough that a negative isothermal compressibility can be deduced, if an unfortunate selection of truncation distance is made. This behavior is representative for all simulations described in this section. Numerical integration of the non-extended RDFs is clearly not a robust approach even for rather simple systems like these.

The Verlet method, as described in Chapter 5, was applied to extend and integrate the TCFs obtained from the simulations. The matching distance parameter R (see Equation (5.9)) was selected using the principles described in Section 5.5. This generally yielded values of R around 1.5σ, which roughly corresponds to the location of the minimum after the first peak of $g(r)$ (see Figure 6.2(a)). As will be seen below, integration of the Verlet-extended TCFs yielded isothermal compressibilities in overall good agreement with the Mecke *et al.* (1996) and Gross and Vrabec (2006) EOS. A remarkable conclusion is that applying an approximation for the long-range DCF allows the compressibility to be predicted using a rather limited part of the sampled $g(r)$. This might open up for the possibility to predict thermodynamic

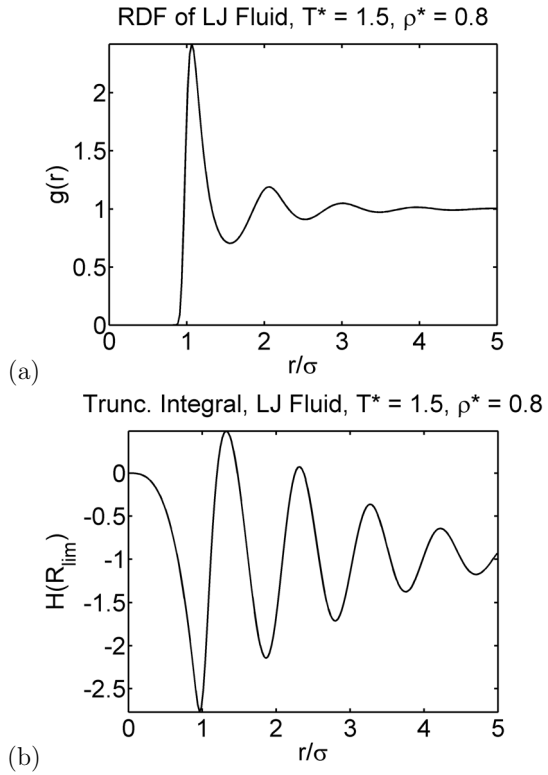


Figure 6.2: (a) RDF obtained from simulation of the pure LJ fluid at a reduced temperature of 1.5 and reduced density of 0.8 and (b) the truncated integral of the same RDF as a function of the upper integration limit R_{lim}. The integral does not converge within the sampling limit which is approximately 5σ and corresponds to half the dimension of the simulation box.

properties efficiently by simulations of small systems.

Lennard-Jones Fluid Figure 6.3(a)–(d) compares the isothermal compressibilities obtained from the Verlet method with those obtained from the Mecke *et al.* (1996) EOS. At all four temperatures, the Verlet results were qualitatively consistent with the EOS, and the discrepancies were typically 1-5%. The situation was not as good at the two lowest temperatures where the differences were as high as 6%. The greatest disagreement was seen when $T^* = 1.5$ and ρ^* is 0.3 or 0.4, which are the state points closest to the critical point at $\rho_c^* = 0.304$ and $T_c^* = 1.316$ (Smit, 1992). At those conditions, the differences were 7% and 8%, respectively. This disagreement is not surprising considering that the quantity $1 - C$ can be very small in this region. Thus, the compressibility as given by Equation (6.5) will be more sensitive to the possible inaccuracies introduced by enforcing an approximation for the DCF for $r > R$. In conclusion, the Verlet method as employed here seems to be

best suited for systems at liquid density, which roughly means $\rho > 2\rho_c$.

The disagreement seen in the near-critical region is not a severe limitation of the method since molecular simulation is most appropriately used to model condensed systems. For systems at near-critical or low densities, scaling laws and virial expansions might be more appropriate approaches as they are simpler than simulations, but still accurate under those conditions.

For $\rho^* = 0.7$ and $\rho^* = 0.8$, the agreement with the EOS was better at higher temperature (1% at $T^* = 2.5$) than at lower temperature (6% at $T^* = 0.85$). The temperature effects on the accuracy can be understood considering Equation (5.44) from which it is clear that the tail contribution to the DCFI increases in magnitude with decreasing temperature. It is thus likely that the presumably small error introduced by forcing the tail to follow Equation (5.44) resulted in more pronounced errors in κ_T at low temperatures. It is also possible that derivatives of the Mecke *et al.* (1996) EOS were less accurate at lower temperature since at those conditions, the EOS did not reproduce the simulation pressures as well as it did at higher temperature. Nevertheless, though low temperatures seemed to offer more of a challenge, the results obtained under those conditions are still considered to be satisfactory.

Statistical errors in the results were estimated for a few representative state points using the blocking method, considering blocks of 100 ps to be independent (Allen and Tildesley, 1987). The standard error in κ_T was found to be less than 0.5% of κ_T itself, indicating that the calculations were well converged.

Stockmayer Fluid The isothermal compressibilities obtained from the simulations of the Stockmayer fluid with $\mu^{*2} = 1$ are compared with the Gross and Vrabec (2006) EOS in Figures 6.4(a)–(d). At low or near-critical densities ($\rho < 2\rho_c$, $T^* = 1.5$), the simulation results were in poor agreement with the EOS with discrepancies up to 40%, similarly to the situation for the LJ fluid. This is not shown since focus is on the region of liquid densities, $0.70 \leq \rho^* \leq 0.95$, where the lower bound approximately corresponds to $2\rho_c^*$ ($\rho_c^* = 0.317$ (Gross and Vrabec, 2006)).

Under these conditions, the isothermal compressibilities obtained from the Verlet method were at all temperatures within 3-4% of those obtained from the Gross and Vrabec (2006) EOS. This is a good agreement, considering that the pressures typically disagreed by 0.5-2.5%. Consistent with the LJ results, the agreement improved with increasing temperature. The error was for instance around 3% when $T^* = 1.0$, while it was less than 1% when $T^* = 2.5$. There was no tendency for the agreement to depend on density as long as $\rho^* > 2\rho_c^*$.

The results from the simulations using $\mu^{*2} = 3$ are shown in Figure 6.5(a)–(c). The compressibilities obtained from most of the simulation agreed with the ones obtained from the Gross and Vrabec (2006) EOS to within 6%. For a few simulations, the disagreement is slightly higher namely 8–11%. This is worse than for $\mu^{*2} = 1$, but still fairly acceptable considering that the agreement of pressures also was worse for the higher dipole moment. Just as for the lower dipole moment and for the LJ fluid, the agreement improved with increasing temperature.

The results for the Stockmayer fluid were worse at the higher dipole moment. One possible reason for this is that the simplified OZ equation (Equation (5.8)) deteriorates due to the stronger anisotropic forces, as discussed above. Since satisfactory results nevertheless are obtained for $\mu^{*2} = 3$, the simplified treatment of

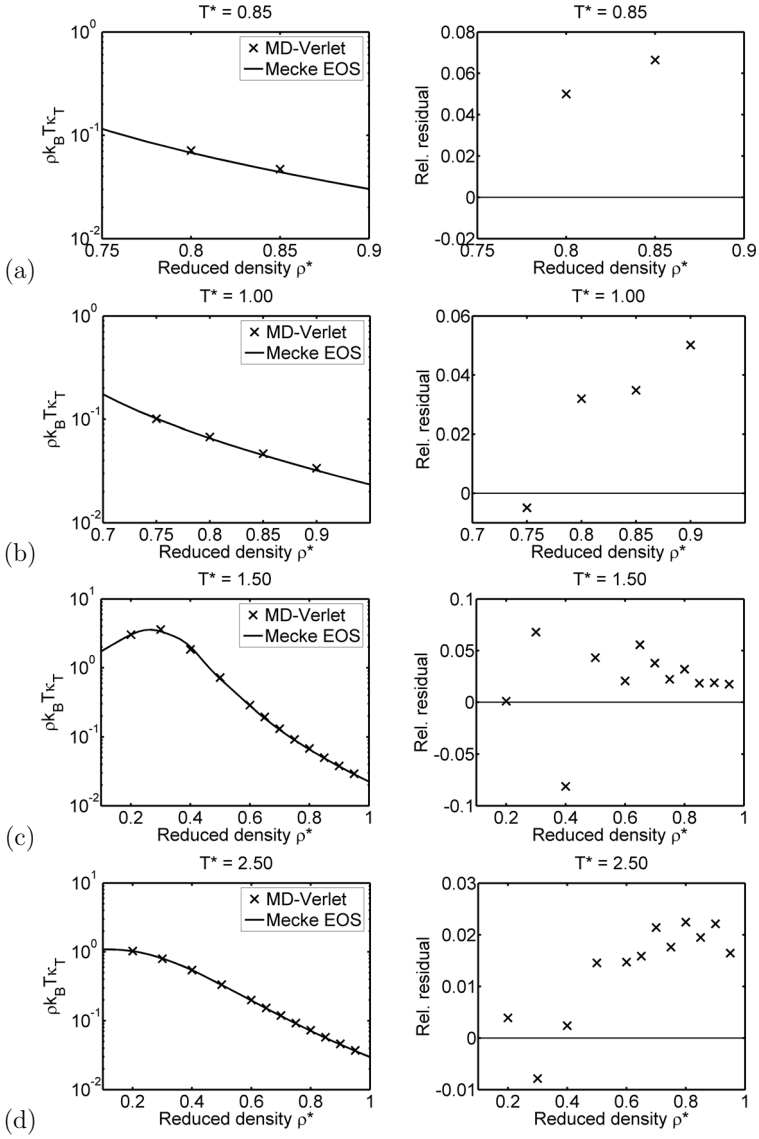


Figure 6.3: Results (left) and relative residuals (right) from calculations of $\rho k_{\mathrm{B}} T \kappa_T$ for the pure LJ fluid at the reduced temperatures (a) $T^* = 0.85$, (b) $T^* = 1.0$, (c) $T^* = 1.5$ and (d) $T^* = 2.5$. The values derived from the Mecke *et al.* (1996) EOS are also shown.

the anisotropic interactions is apparently adequate for the systems described in this section.

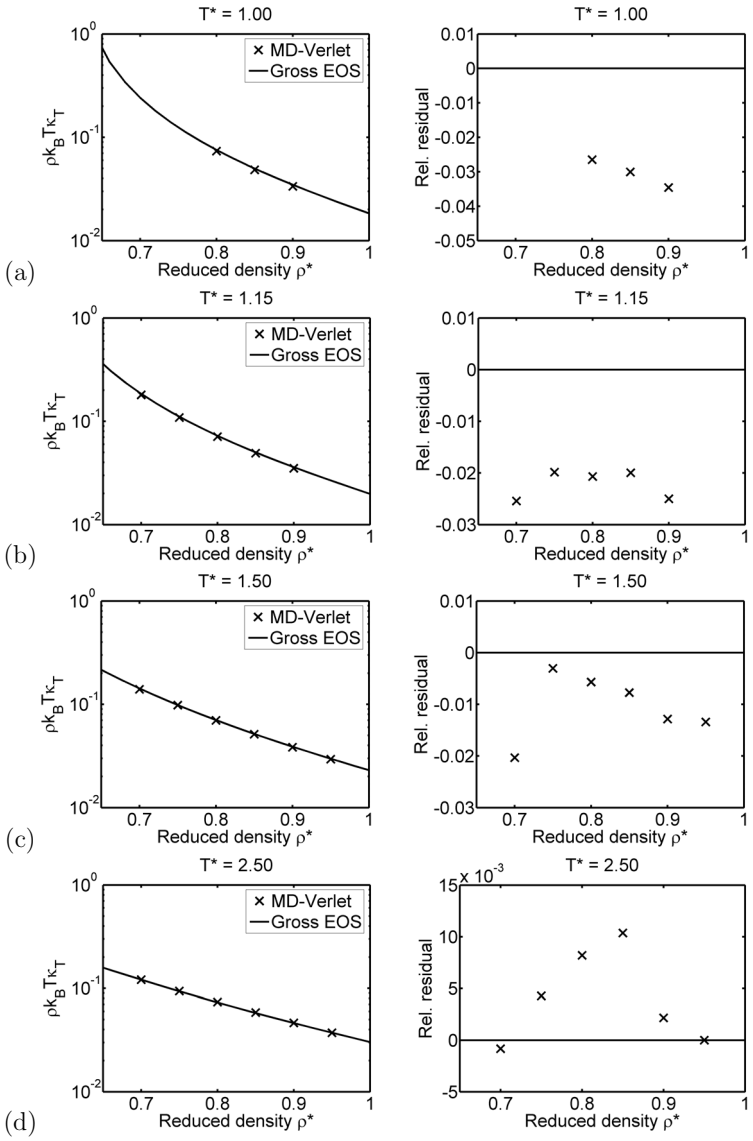


Figure 6.4: Results (left) and relative residuals (right) from calculations of $\rho k_{\mathrm{B}} T \kappa_T$ for the pure Stockmayer fluid with $\mu^{*2} = 1$ at the reduced temperatures (a) $T^* = 1.00$, (b) $T^* = 1.15$, (c) $T^* = 1.5$ and (d) $T^* = 2.5$. The values derived from the Gross and Vrabec (2006) EOS are also shown.

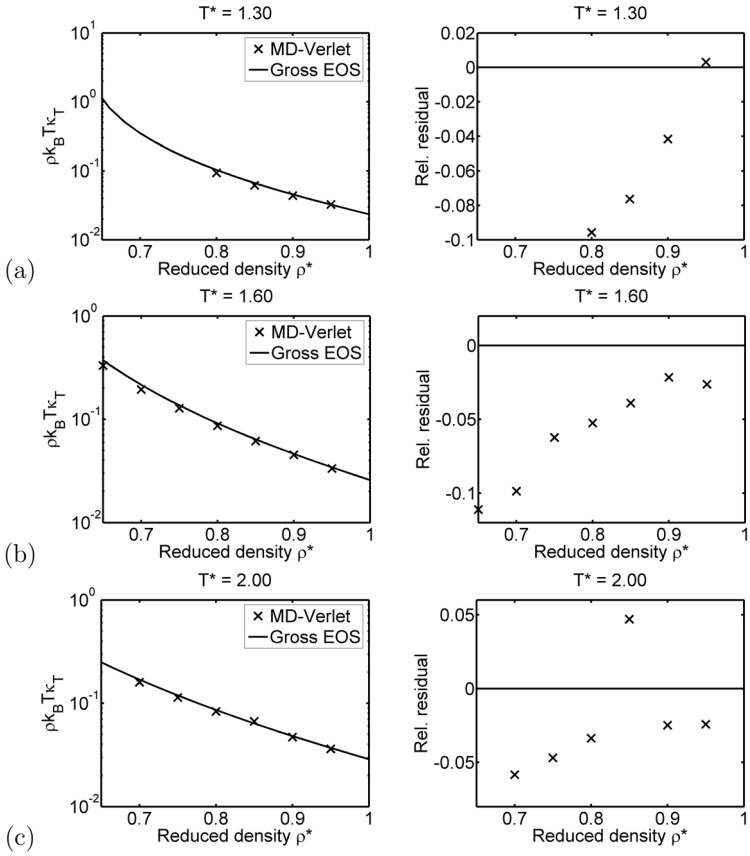


Figure 6.5: Results (left) and relative residuals (right) from calculations of $\rho k_{\mathrm{B}} T \kappa_T$ for the pure Stockmayer fluid with $\mu^{*2} = 3$ at the reduced temperatures (a) $T^* = 1.00$, (b) $T^* = 1.15$, (c) $T^* = 1.5$ and (d) $T^* = 2.5$. The values derived from the Gross and Vrabec (2006) EOS are also shown.

6.2 Lennard-Jones/Stockmayer Mixtures

The study of the pure LJ and Stockmayer fluids of the previous section is naturally extended to a study of LJ/Stockmayer mixtures. Such mixtures includes "Stockmayer" atoms with dipole moment of magnitude μ and "LJ" atoms with zero dipole moment. LJ-LJ and LJ-Stockmayer interactions thus follow the LJ potential (Equation (6.2)), while Stockmayer-Stockmayer interactions include dipole-dipole interactions (Equation (6.3)) as well. The LJ and Stockmayer components are here referred to by subscripts 1 and 2, respectively.

The extension to mixtures allows for prediction of mixture properties which are more relevant for applications than pure-component properties. This is due to the fact that prediction of mixture properties is more challenging for conventional property-prediction methods such as group-contribution methods than is prediction of pure-component properties (Poling *et al.*, 2007). Especially appreciable is prediction of the composition derivative of the activity coefficient which can be obtained from Equation (4.12). As described in Section 4.2.3, this property can be used to determine a G^{E} model for the mixture from which the phase behavior can be deduced.

Calculation of TCFIs for mixtures is however also more challenging for the computational methodology of Chapter 5. For mixtures, the three TCFIs H_{11}, H_{12} and H_{22} are calculated. In the expression for the activity coefficient derivative (Equation (4.12)), these integrals enter only through the difference $\Delta H \equiv H_{11} + H_{22} - 2H_{12}$ which needs to be of acceptable accuracy as well as the individual TCFIs.

6.2.1 Simulation details

The simulations of LJ/Stockmayer mixtures were carried out such as the simulations of the pure Stockmayer systems described in Section 6.1.1, using the same values for the LJ parameters, as well as the same simulation program. All simulations comprised 864 particles and were run at a reduced temperature and density of $T^* = 1.15$ and $\rho^* = 0.822$, respectively, and with μ^{*2} for the Stockmayer particles set to 1, 2 or 3. These particular state conditions were chosen since most previous simulations of LJ/Stockmayer mixtures seem to have been carried out at these conditions (de Leeuw *et al.*, 1990). They are furthermore liquid states and previous results indicate that the two components are fully miscible at these dipole moments (de Leeuw *et al.*, 1990). For each of the three dipole moments studied, simulations were carried out at the compositions $x_1 = 0.125$, 0.25, 0.375, 0.5, 0.625, 0.75 and 0.875, where x_1 denotes the fraction of LJ particles.

The simulations were equilibrated for at least 100 ps, and the total production time for each composition was 19.4 ns. RDFs for like-like and like-unlike species were sampled using a bin width of 0.03σ. The production periods were significantly longer than those used with the corresponding pure systems since properties expressed as differences of TCFIs converged more slowly than the TCFIs themselves. Estimates of statistical uncertainties were based on eight independent simulations which were started with different starting velocities.

6.2.2 Results

Ideally, the TCFIs obtained from integration of the extended TCFs should be validated by calculating the "target properties" isothermal compressibility, partial molecular volumes and activity coefficient derivatives (Equations (4.10)–(4.12)) and compare these with property values obtained from alternative routes or from correlations of previous simulations of the same systems.

The Gross and Vrabec (2006) EOS allows for mixtures of fluid particles with different dipole moments and could thus, in principle, be used to obtain the target properties for LJ/Stockmayer mixtures. The average pressures and dipole-dipole energies obtained from simulations did however not agree satisfactorily with the EOS. The agreement in pressure was overall reasonable but deteriorated with increasing μ^2 and increasing x_2 (fraction of Stockmayer particles). The discrepancies were as large as 15% at Stockmayer-rich compositions with $\mu^{*2} = 3$. The agreement in dipole-dipole energy did as well deteriorate as μ^2 increased. This is not surprising since the Gross and Vrabec (2006) EOS is based on perturbation theory. The agreement did also deteriorate with decreasing x_2. At LJ-rich compositions with $\mu^{*2} = 3$, the discrepancies were as large as 50%. In particular since the property agreement deteriorated systematically as the composition was varied, the EOS would probably not reproduce composition derivatives accurately since inaccuracies usually are aggravated by differentiation. The Gross and Vrabec (2006) EOS was therefore not employed to validate partial molecular volumes or activity coefficient derivatives obtained from the extended TCFs. The EOS was nevertheless used for validation of the isothermal compressibility since this property is based on differentiation with respect to density rather than composition.

Since no other EOS for LJ/Stockmayer mixtures seems to exist, the validation of composition-derivative properties was instead approached by considering the relations

$$\frac{1}{k_\mathrm{B}T}\left(\frac{\partial^2 A^\mathrm{E}}{\partial x_1^2}\right)_{N,V,T} = 2C_{12} - C_{11} - C_{22} \equiv -\Delta C \tag{6.6}$$

$$\frac{1}{\rho k_\mathrm{B}T}\left(\frac{\partial P}{\partial x_2}\right)_{N,V,T} = x_1 C_{11} - x_2 C_{22} + (x_2 - x_1)C_{12} \tag{6.7}$$

where A^E, x_1, N, V and P respectively denote the excess Helmholtz energy per molecule, the number fraction of component 1 (LJ), the total number of particles, the total volume and the pressure. C_{ij} denotes the DCFI for the pair ij and is defined by

$$C_{ij} \equiv 4\pi\rho \int r^2 c_{ij}(r) dr \tag{6.8}$$

where $c_{ij}(r)$ denotes the DCF for the pair ij. Equations (6.6) and (6.7), which are derived in Appendix E, are here expressed in DCFIs rather than TCFIs, as the latter would lead to more complicated expressions. The properties were obtained by integration of the DCFs obtained from the Verlet method and validated against values obtained from alternative routes.

In order to validate $(\partial P/\partial x_2)_{N,V,T}$, the cubic polynomial

$$P_\mathrm{model} = a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^3 \tag{6.9}$$

was fitted to pressures obtained from the simulations. The property of Equation (6.7) was determined by analytical differentiation of the fitted polynomial. This was carried out for each of the three values of μ^2 studied.

Values of ΔC were obtained from the Helmholtz energy model by de Leeuw *et al.* (1990) which is an expression fitted to average dipole-dipole energies obtained from simulations of LJ/Stockmayer mixtures. The model as well as how it was applied here is described in a separate paragraph below.

The relevance of the derivative property ΔC can be realized by writing the activity coefficient derivative as

$$\left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P,N_2} = \frac{1}{k_\mathrm{B}T}\left[x_2\left(\frac{\partial^2 A^\mathrm{E}}{\partial x_1^2}\right)_{N,V,T} - \frac{\rho(\bar{v}_1 - \bar{v}_2)^2}{\kappa_T}\right] \tag{6.10}$$

where γ_1, N_1 and $\bar{v}_1$ respectively denote activity coefficient, particle number and molecular volume of component 1 (LJ). The equation is derived in Appendix E. For strongly non-ideal liquid mixtures, the major contribution comes from the first term within the square brackets. For such systems, the accuracy of the predicted ΔC indicates how accurately one can expect to obtain the activity coefficient derivative, which is the property of greatest interest for applications.

An important remark is that the three "validation properties", $\rho k_\mathrm{B}T\kappa_T$, $(\partial P/\partial x_2)_{N,V,T}$ and ΔC are independent, linear functions of the three DCFIs[1]. If these properties are accurately obtained, it follows that also the three DCFIs and TCFIs are accurate.

The de Leeuw Helmholtz energy model The de Leeuw *et al.* (1990) Helmholtz energy model is based on equating the reduced average dipole-dipole interaction energy per molecule, $U^{(\mathrm{dd})*}$, with the Padé expression

$$U^{(\mathrm{dd})^*}(\mu^{*2}, x_2) = \frac{-Ax_2^2\mu^{*4}}{1 + C(x_2)\mu^{*2}} \tag{6.11}$$

where x_2 denotes the fraction of Stockmayer particles and where A is related to properties of the pure LJ fluid under similar state conditions. In the study of de Leeuw *et al.* (1990), A was estimated from simulations of the pure LJ fluid, and the function $C(x_2)$ was approximated by a linear function, whose parameters were fitted to reproduce values of $U^{(\mathrm{dd})*}$ obtained from simulations. The molecular excess Helmholtz energy is obtained by thermodynamic integration of Equation 6.11 with respect to μ^{*2}, which results in

$$\begin{aligned} A^\mathrm{E} = &-\frac{Ax_2^2\left[\mu^{*2}C(x_2) - \ln\left(1 + \mu^{*2}C(x_2)\right)\right]}{\left[C(x_2)\right]^2} \\ &+\frac{A\left[\mu^{*2}C(1) - \ln\left(1 + \mu^{*2}C(1)\right)\right]}{\left[C(1)\right]^2} \end{aligned} \tag{6.12}$$

Although the parameters supplied by de Leeuw *et al.* (1990) reproduce dipole-dipole interaction energies obtained from the present simulations reasonably well, a better fit could be achieved by revising the parameterization. For this purpose, relatively

[1]It is actually $(\rho k_\mathrm{B}T\kappa_T)^{-1}$ which is linear function of the DCFIs as shown by O'Connell (1971b).

short simulations of the LJ/Stockmayer mixture were carried out with equilibration and production periods of 100 and 800 ps, respectively, and with x_2 set to 0.125, 0.25, 0.375, 0.5, 0.625, 0.75 or 0.875, and μ^{*2} set to 0.5, 1.0, 1.5, 2.0, 2.5 or 3.0. Equation 6.11 was adjusted to reproduce values of $U^{(\mathrm{dd})*}$ obtained from the simulations. The parameter A was as well fitted. The function $C(x_2)$ was parameterized as a polynomial. A quadratic polynomial yielded a better fit than a linear. A cubic polynomial did however not improve the fit, and thus, the quadratic polynomial was retained. The original parameters reported by de Leeuw *et al.* (1990) were $A = 1.70$ and $C(x_2) = 0.876x_2 - 0.134$ while the parameters obtained here were $A = 1.58$ and $C(x_2) = -0.177x_2^2 + 1.052x_2 - 0.203$.

The derivative property ΔC was in accordance with Equation (6.6) obtained by twice differentiating the excess Helmholtz expression of Equation (6.12).

Consistency of ΔC Figures 6.6(a)–(c) compares values of $-\Delta C$ obtained by the Verlet method with those obtained from the de Leeuw *et al.* (1990) Helmholtz energy correlation. At all three dipole moments, the results from the extension method were in good agreement with the correlation using the new parameterization.

Although the simulations at each dipole moment had the exact same production times, the relative statistical uncertainties were largest with $\mu^{*2} = 1$ and the agreement with the correlation was slightly worse than for $\mu^{*2} = 2$ or 3. This deviation can be attributed to that the system becomes more non-ideal with increasing dipole moment. Previous studies suggest that fluctuation solution theory (FST) modeling is best suited for strongly non-ideal systems (Christensen *et al.*, 2007c) where the values of ΔC are large in magnitude. At nearly ideal conditions, Equation (6.6) becomes a difference between nearly equal numbers with uncertainties, and thus yields results of lower accuracy and precision.

Values of $-\Delta C$ from the Verlet method overestimated those from the correlation when either species was dilute. This was pronounced for $\mu^{*2} = 1$ (Figure 6.6(a)) where differences were greatest at LJ-rich and Stockmayer-rich compositions. For $\mu^{*2} = 2$ and 3, the agreement was better and the deviations were only seen at Stockmayer-rich compositions. Two factors may offer partial explanations of this. At Stockmayer-rich compositions, the dipole-dipole interactions are more effectively screened than they are at compositions of lower Stockmayer content. When x_2 increases, the Stockmayer particles thus behave more like LJ particles, and the mixture resembles more an ideal mixture for which FST is less accurate. At LJ-rich compositions, the dipole-dipole interactions are screened to only a lesser extent resulting in a more non-ideal system than at Stockmayer-rich compositions. A second factor is that at compositions where one component is dilute, the simulation includes only a small number of atoms of this component. This probably leads to less efficient sampling of the RDFs involving this component and results of higher susceptibility to finite-size effects such as those discussed by Salacuse *et al.* (1996).

The accuracy of calculated like-like TCFIs was also seen to deteriorate as the corresponding component became dilute in the study of Christensen *et al.* (2007a). The authors attributed this to the lower sampling efficiency for the dilute species.

Consistency of $(\partial P/\partial x_2)_{N,V,T}$ Overall, the pressure composition derivatives obtained by the Verlet method were in good agreement with those obtained from the

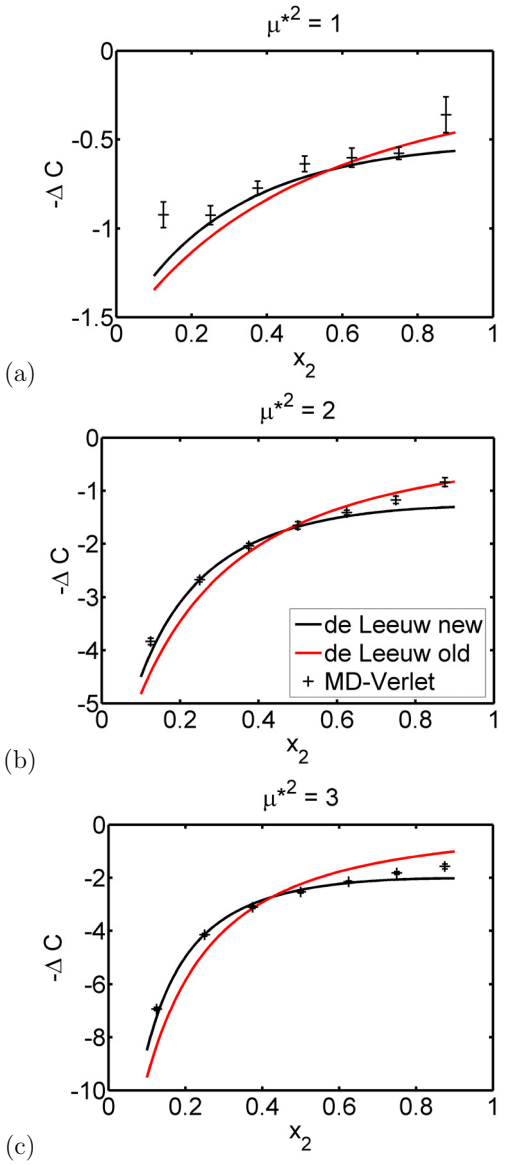


Figure 6.6: Values of $-\Delta C$ for LJ/Stockmayer mixtures vs. the fraction of Stockmayer particles x_2 for dipole moments of (a) $\mu^{*2} = 1$, (b) $\mu^{*2} = 2$ and (c) $\mu^{*2} = 3$, derived from the de Leeuw Helmholtz energy model using either the original parameters (de Leeuw *et al.*, 1990) (red line) or the refitted parameters (black line), compared with the Verlet method results (crosses). Standard error estimates were for each data point based on eight simulations which were started from different initial velocities.

polynomial fit of Equation (6.9) (Figures 6.7(a)–(c)). The errors were around 1-4.5 % or less for most of the systems. The disagreement was largest for $x_2 = 0.125$ or 0.875, i.e. when either of the species was dilute. At all compositions, the differences decreased with increasing dipole moment, i.e. as the mixtures became more non-ideal. This is consistent with the above discussion, and the trends can be attributed to the same factors.

Consistency of $\rho k_B T \kappa_T$ The isothermal compressibilities obtained from the Verlet method are in Figures 6.8(a)–(c) compared with the Gross and Vrabec (2006) EOS. For $\mu^{*2} = 1$ (Figure 6.8(a)), the Verlet values agrees very well with the EOS, with differences around 1–1.5%. For the higher dipole moments, the agreement is still good when x_2 is small but deteriorates as x_2 increases (Figure 6.8(b)–(c)). This is more pronounced for $\mu^{*2} = 3$ where the discrepancies becomes as large as 11% at Stockmayer-rich composition.

As discussed above, the Gross and Vrabec (2006) EOS did not reproduce simulation pressures fully satisfactorily as the agreement worsened with increasing μ^{*2} and x_2. This is probably the reason why the exact same trends to occur for the isothermal compressibility.

6.2.3 Comparison with Previous Integration Approaches

It is from the previous section clear that the Verlet method as applied here predicts derivative properties that are in overall good consistency with the benchmark values. It is relevant to investigate whether alternative approaches to compute TCFIs from simulation achieve the same accuracy as the Verlet method. The truncation method (Weerasinghe and Smith, 2003) and the Hess method (Hess and van der Vegt, 2009) were therefore employed to calculate the validation properties $\rho k_B T \kappa_T$, $(\partial P/\partial x_2)_{N,V,T}$ and ΔC as defined above.

As discussed in Section 4.2.2, simple truncation requires averaging the function $H(R_{\text{lim}})$ over a specific interval, where $H(R_{\text{lim}})$ is the numerical TCFI as a function of the upper integration limit R_{lim}. It is not obvious how to choose this interval and no robust approach seems to exist. Here, the TCFIs were averaged with R_{lim} in the interval $[2\sigma, 3\sigma]$, where σ roughly corresponded to the oscillation period of $h_{ij}(r)$.

With the Hess method, the scaling factor α_{ij} was evaluated using Equation (4.16), with $R = 4\sigma$. The RDFs were rescaled, but the integrals did still not converge within the sampled range. The integration of the rescaled RDFs was thus carried out by the same principles as with the truncation method, but using larger truncation radii. This was based on the idea that the corrected RDFs are more reliable at large separations than are uncorrected ones. The TCFIs were averaged using truncation radii in the interval $[3.5\sigma, 4.5\sigma]$.

The truncation and Hess methods reproduced qualitatively well the values of $-\Delta C$ from the Helmholtz energy model (Figures 6.9(a-c) and Table 6.2). For truncation, the discrepancies were 20-30 % when $\mu^{*2} = 1$ and 10-20 % when $\mu^{*2} = 2$ and 3 and became much worse when either component was dilute. These discrepancies were consistently larger than for the Verlet method results. The results of the Hess method were in better overall agreement with the Helmholtz energy correlation than the ones from simple truncation, for $\mu^{*2} = 2$ and 3. The Verlet method however

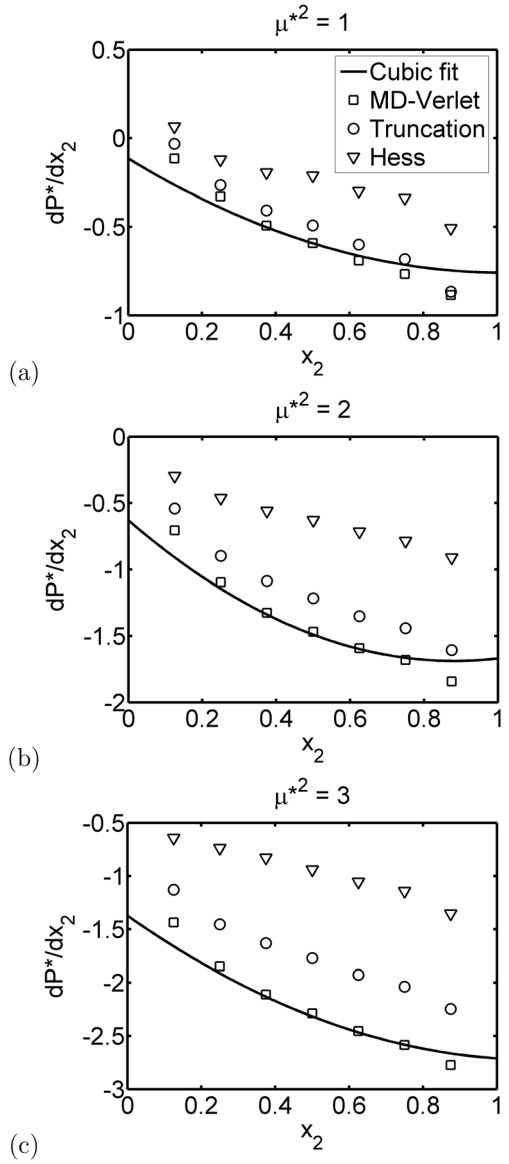


Figure 6.7: Values of $(\partial P/\partial x_2)_{N,V,T}$ for LJ/Stockmayer mixtures vs. the fraction of Stockmayer particles x_2 for dipole moments of (a) $\mu^{*2} = 1$, (b) $\mu^{*2} = 2$ and (c) $\mu^{*2} = 3$, derived from fitting a cubic polynomial to the average pressures (line), compared with the results from either the Verlet (squares), truncation (circles) or Hess (triangles) methods.

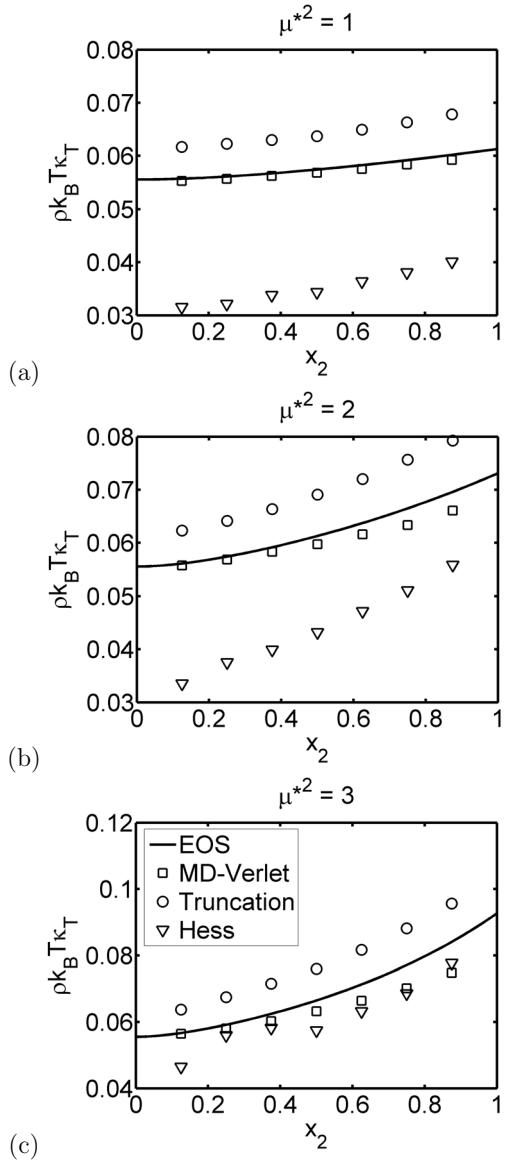


Figure 6.8: Values of $\rho k_{\mathrm{B}} T \kappa_T$ for LJ/Stockmayer mixtures vs. the fraction of Stockmayer particles x_2 for dipole moments of (a) $\mu^{*2} = 1$, (b) $\mu^{*2} = 2$ and (c) $\mu^{*2} = 3$, derived from the EOS of Gross and Vrabec (2006) (line), compared with the results from either the Verlet (squares), truncation (circles) or Hess (triangles) methods.

performed either equally well or significantly better than the Hess method, except at $x_2 = 0.875$.

For the pressure composition derivative (Figures 6.7), both simple truncation and the Hess method results were in poor agreement with the results from the polynomial fit. The discrepancies were 10-30 % for simple truncation and 30-60 % for the Hess method. This indicated that these methods did not yield accurate individual TCFIs, and that the reasonable agreement for $-\Delta C$ probably originated from cancellation of the errors in the individual TCFIs.

The values obtained for the isothermal compressibility (Figure 6.8(a)–(c)) further confirms the failure of simple truncation and the Hess method. Truncation overestimates the compressibility by 10–15% while the Hess method underestimates it by 10–40%.

Also notable is that the Verlet method yielded more statistically precise results for $-\Delta C$ than the truncation method (Table 6.2) since it used less information from the sampled RDFs. Truncation, as employed here, yielded similarly more precise results than the Hess method.

As stated in Section 4.2.2, the truncation and Hess methods are sensitive to the choice of truncation radii. It is therefore possible that better results would have been obtained when using other values than applied here. It is also likely that these methods would perform better if significantly larger systems would have been considered so that the TCFs could be reliably integrated numerically over a longer range.

6.2.4 Concluding Remarks

The Verlet method, as described in Chapter 5, yielded accurate results for LJ/Stockmayer mixtures as indicated by comparisons with benchmark values. In particular, accurate individual correlation function integrals were obtained. The method achieved furthermore a better accuracy than simpler integration approaches. Caution is however advised when the studied system is nearly ideal, or when one component is dilute is present in a fraction less than approximately 15%.

6.3 Pure Molecular Fluids

Molecular models resembling "real" substances most often involve multiple interaction sites. The successful applications to atomic fluids do not necessarily imply that the Verlet method of Chapter 5 is accurate for molecular fluids since these are treated in a more approximate way. The focus of this section and Section 6.4 is therefore to evaluate the performance for simulations of molecular fluids with atom-atom interactions.

The treatment of molecular fluids is more approximate for two reasons. Firstly, the non-spherical geometry of molecules leads to anisotropic interactions that are different from those present in the Stockmayer fluids. These interactions might challenge the approximate treatment of the OZ equation which assumes decoupling of the isotropic and anisotropic correlations (Equation (5.8)). In particular, the short-ranged repulsive forces between non-spherical molecules seem to affect the

Table 6.2: Results for ΔC obtained from the three different approaches to evaluate TCFIs, compared with the values derived from the de Leeuw *et al.* (1990) Helmholtz energy model using the new parameters. Standard errors estimates were for each value based on eight simulations which were started from different initial velocities.

μ^{*2}	x_2	Verlet	Truncation	Hess	de Leeuw
1	0.125	0.92 ± 0.07	0.6 ± 0.2	1.2 ± 0.3	1.20
	0.250	0.93 ± 0.05	0.73 ± 0.09	1.0 ± 0.2	0.97
	0.375	0.77 ± 0.04	0.62 ± 0.09	1.0 ± 0.2	0.81
	0.500	0.64 ± 0.04	0.5 ± 0.1	0.7 ± 0.1	0.71
	0.625	0.60 ± 0.05	0.5 ± 0.1	0.8 ± 0.1	0.64
	0.750	0.58 ± 0.03	0.4 ± 0.1	1.1 ± 0.1	0.60
	0.875	0.4 ± 0.1	0.1 ± 0.2	1.1 ± 0.4	0.57
2	0.125	3.84 ± 0.06	3.3 ± 0.1	3.8 ± 0.2	4.05
	0.250	2.68 ± 0.04	2.39 ± 0.06	2.95 ± 0.09	2.68
	0.375	2.04 ± 0.05	1.78 ± 0.09	2.04 ± 0.07	2.03
	0.500	1.65 ± 0.07	1.47 ± 0.09	1.79 ± 0.08	1.68
	0.625	1.41 ± 0.04	1.25 ± 0.10	1.7 ± 0.1	1.48
	0.750	1.17 ± 0.07	1.0 ± 0.1	1.6 ± 0.2	1.37
	0.875	0.84 ± 0.08	0.4 ± 0.2	1.0 ± 0.4	1.31
3	0.125	6.93 ± 0.06	5.98 ± 0.09	6.9 ± 0.3	7.24
	0.250	4.15 ± 0.05	3.55 ± 0.09	4.2 ± 0.2	4.13
	0.375	3.11 ± 0.05	2.67 ± 0.09	3.2 ± 0.2	2.98
	0.500	2.53 ± 0.05	2.21 ± 0.08	2.6 ± 0.2	2.46
	0.625	2.15 ± 0.03	1.93 ± 0.07	2.6 ± 0.2	2.20
	0.750	1.82 ± 0.05	1.6 ± 0.1	2.4 ± 0.1	2.07
	0.875	1.58 ± 0.09	1.3 ± 0.2	1.9 ± 0.3	2.02

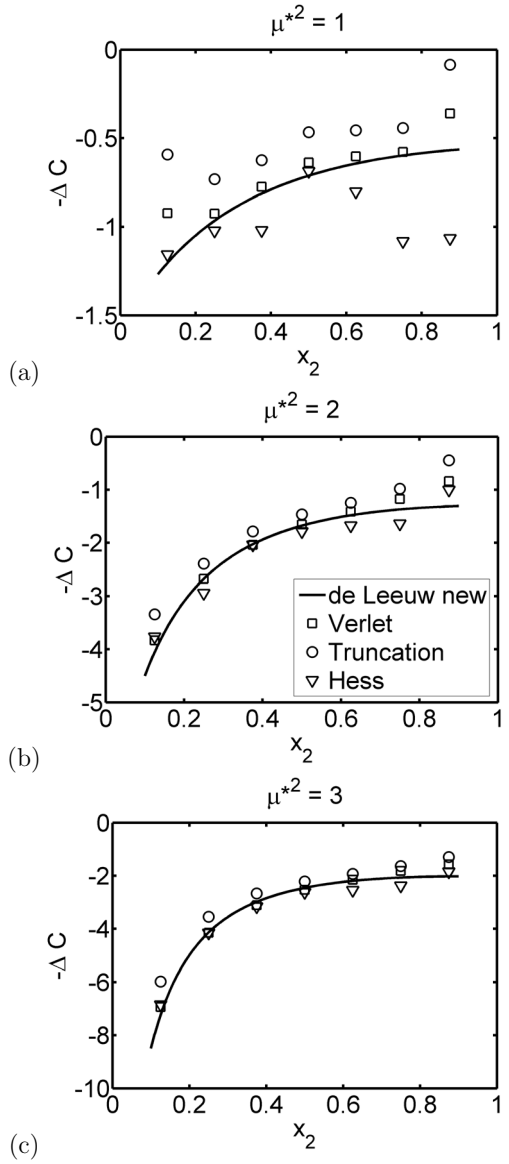


Figure 6.9: Values of $-\Delta C$ for LJ/Stockmayer mixtures with (a) $\mu^{*2} = 1$, (b) $\mu^{*2} = 2$ and (c) $\mu^{*2} = 3$, derived from the de Leeuw Helmholtz energy model using the refitted parameters (solid line), compared with the results from either the Verlet (squares), truncation (circles) or Hess (triangles) methods.

orientation-averaged RDF to a greater extent than dipole-dipole interactions do, as shown by Monte Carlo simulation by Wang *et al.* (1973). Secondly, the DCF tail approximation of Equation (5.44) retains only the r^{-6} term in the power series expansion of the angle-averaged pair potential. As shown in Appendix D, this term is independent of the molecular geometries as it is determined by the atom types only. The neglected terms of order r^{-8} and higher do however depend on the molecular geometries and become more significant as the molecules become more non-spherical.

The Verlet method was tested on simulations of pure molecular liquids. The studied liquids were ethane, (n-)butane, (n-)hexane, water and 2-propanol. The CHARMM force field (MacKerell Jr. *et al.*, 1998) was used to model the compounds. Linear alkanes were considered due to their relatively simple structure and since the chain length serves as an order parameter for how non-spherical the molecule is. Pure water and 2-propanol were considered in order to test the Verlet method for polar molecular fluids as well.

6.3.1 Simulation details

Simulations of pure ethane, butane, hexane, water and 2-propanol were carried out in the NPT ensemble (particle number, pressure and temperature were constant) using the CHARMM force field (MacKerell Jr. *et al.*, 1998) and the CHARMM-adapted TIP3P model with flexible bonds for water (MacKerell Jr. *et al.*, 1998), at state conditions summarized in Table 6.3. Each simulation comprised 1000 molecules except for the pure water simulation, in which 2000 molecules were considered. The velocity Verlet algorithm with a 1 fs time step was employed to integrate the equations of motion, periodic boundary conditions were employed in the x, y and z directions, and electrostatic forces were evaluated using the particle mesh Ewald method with a grid spacing smaller than 1 Å. Temperature and pressure were respectively controlled using the Langevin thermostat algorithm with a damping constant of 5 ps^{-1} and the Langevin piston algorithm with a period of 200 fs and a decay constant of 500 fs. Coordinates were sampled each 500 fs. For the pure water and 2-propanol simulations, LJ forces were evaluated using a 12 Å cutoff, a 10 Å switching distance, and using a pair list with an outer radius of 14 Å. The alkane simulations used a 15 Å cutoff, a 13 Å switching distance and a pair list outer radius of 17 Å were used. The reason for this is that some of the simulations were carried out at state conditions where the fluid is rather compressible. For such systems, long-range correlations play a more important role than for dense liquids, and the results are more likely to be sensitive to the truncation of LJ forces.

The systems were equilibrated for at least 200 ps, and the production periods were 8 ns, 10 ns and 4 ns for the alkane, 2-propanol and water systems, respectively.

6.3.2 Results

As explained previously (Section 6.1.2), the most relevant derivative property for pure fluids is the isothermal compressibility. The values of this property obtained from the Verlet method were compared to those obtained via the fluctuation formula,

Table 6.3: Lists temperature T and pressure P for the simulations of pure molecular fluids, and densities calculated from the simulations ρ_{MD}. Experimental critical temperatures $T_{\mathrm{c,exp}}$ are also listed for each fluid.

	ID	$T_{\mathrm{c,exp}}$	T [K]	P [atm]	ρ_{MD} [g/dm^3]
Ethane	EthA	296	380	600	403
	EthB		305	225	399
	EthC		260	100	438
	EthD		180	100	561
	EthE		120	100	636
Butane	ButA	425	500	350	426
	ButB		425	140	424
	ButC		340	100	531
	ButD		260	100	623
	ButE		180	100	704
Hexane	HexA	508	610	300	424
	HexB		508	120	437
	HexC		400	100	561
	HexD		300	100	658
	HexE		200	100	746
Water		647	323	1	1018
2-propanol		509	298	1	781

given by (Allen and Tildesley, 1987)

$$\kappa_T = \frac{\langle V^2 \rangle_{NPT} - \langle V \rangle^2_{NPT}}{k_{\mathrm{B}} T \langle V \rangle_{NPT}} \tag{6.13}$$

where V denotes simulation box volume, and $\langle \cdot \rangle_{NPT}$ denotes isothermal-isobaric (NPT) ensemble average. The results are listed in Table 6.4.

For the alkane systems, the Verlet method agreed well with the fluctuation formula with discrepancies of 2-9 % in general (Table 6.4). These discrepancies seemed uncorrelated with temperature and density. Similarly to the LJ and Stockmayer fluids (Section 6.1.2), good agreement was seen at high, supercritical temperatures, but also for some of the simulations at low temperature, i.e. EthE, ButE, HexD and HexE. The deviations were slightly larger than for the pure LJ and Stockmayer fluids in the dense region, which is expected due to that the molecules are more non-spherical. Nevertheless, the Verlet results were still of acceptable accuracy. The deviations did not seem to increase as the molecules became less spherical. For the water simulation, the agreement was 7.8 % which seems good, considering the strongly anisotropic character of this fluid. The agreement for 2-propanol was 1.4 %, which is very good.

From Table 6.4, it is also apparent that larger molecules require larger matching distances R. Lower temperatures also seem to require larger values of R, which is most apparent for the hexane simulations. For HexE, the R used is only slightly smaller than half the simulation box dimension. For molecules that are larger than hexane and studied at similar temperatures, one might need to consider systems composed of more than 1000 molecules.

6.3.3 Comparison with the Truncation Method

For comparison with the Verlet method results, the TCFs obtained from the simulated systems were also integrated using the truncation approach of Section 4.2.2. The averaging intervals used were 14-18 Å, 16-20 Å, 18-23 Å, 17-21 Å and 13-15 Å for ethane, butane, hexane, 2-propanol and water, respectively. The results for isothermal compressibilities are listed in Table 6.4.

As apparent from the table, the Verlet method generally performed better than truncation, although truncation did better for five of the simulations. For the alkanes, the two methods performed similarly at supercritical temperature. The truncation method however failed completely for EthE, ButC, ButE, HexD and HexE, with deviations from the fluctuation formula exceeding 20 %. For EthE and ButE, integral truncation yielded negative compressibilities which obviously are unphysical.

This demonstrates that truncation is an adequate approach when $g(r)$ is of shorter range than the system size, which is more likely the case at high temperature. Truncation is however unreliable when the oscillations of $g(r)$ still are significant beyond the sampling limit, occurs at low temperature.

For the water simulation, truncation performed just slightly worse than the Verlet method, due to the fact that such a large system is considered for the integral to be sufficiently converged within the sampled range. The opposite was seen for 2-propanol, where truncation failed.

Table 6.4: Results for isothermal compressibility for the simulations of pure molecular fluids obtained via the fluctuation formula, the Verlet method and simple integral truncation. The matching distances R used with the Verlet method are also listed. Standard errors were estimated by the blocking method (Allen and Tildesley, 1987).

	$(\rho k_\mathrm{B} T \kappa_T)_\mathrm{fluc}$	$(\rho k_\mathrm{B} T \kappa_T)_\mathrm{Verlet}$	Dev. [%]	$(\rho k_\mathrm{B} T \kappa_T)_\mathrm{trunc}$	Dev. [%]	R [Å]
EthA	0.183 ± 0.003	0.1934 ± 0.0002	5.5	0.203 ± 0.002	10	9.5
EthB	0.286 ± 0.004	0.303 ± 0.003	6.0	0.298 ± 0.005	4.1	10.4
EthC	0.230 ± 0.003	0.2361 ± 0.0007	2.6	0.239 ± 0.002	3.5	10.2
EthD	0.0411 ± 0.0003	0.0446 ± 0.0001	8.3	0.0461 ± 0.0003	12	12.3
EthE	0.0143 ± 0.0003	0.0141 ± 0.0001	1.8	-0.0235 ± 0.0002	264	11.1
ButA	0.224 ± 0.006	0.2349 ± 0.0005	4.8	0.229 ± 0.003	2.3	11.7
ButB	0.37 ± 0.02	0.383 ± 0.003	2.5	0.357 ± 0.002	4.3	13.3
ButC	0.109 ± 0.001	0.118 ± 0.001	8.6	0.130 ± 0.002	20	14.5
ButD	0.0357 ± 0.0006	0.0391 ± 0.0002	9.4	0.0369 ± 0.0002	3.2	15.0
ButE	0.0131 ± 0.0003	0.0137 ± 0.0002	4.7	-0.020 ± 0.001	249	14.8
HexA	0.225 ± 0.006	0.238 ± 0.001	5.5	0.227 ± 0.002	0.7	16.7
HexB	0.330 ± 0.005	0.342 ± 0.006	3.7	0.311 ± 0.008	5.9	15.5
HexC	0.083 ± 0.003	0.0889 ± 0.0009	7.7	0.085 ± 0.002	3.5	18.2
HexD	0.0274 ± 0.0004	0.0285 ± 0.0003	3.9	0.0200 ± 0.0001	27	27.8
HexE	0.0090 ± 0.0001	0.0093 ± 0.0001	3.1	0.0022 ± 0.0009	76	27.5
Water	0.0772 ± 0.0004	0.0832 ± 0.0001	7.8	0.0849 ± 0.0001	10	8.6
2-propanol	0.0377 ± 0.0006	0.0382 ± 0.0003	1.4	0.0233 ± 0.0009	38	13.1

6.4 Water/Organic Mixtures

As previously stated, the ultimate goal of these efforts is to establish an integration method that reliably predicts activity coefficient derivatives, partial molecular volumes and isothermal compressibilities from simulations of molecular mixtures carried out using an atom-atom interaction model. This section is focused on such applications. Simulations were carried out of three binary mixtures, namely water/acetone, water/methanol and water/t-butanol. The results obtained in this chapter are employed in Chapter 7 in order to determine the water activity in simulations of *Candida antarctica* lipase B in acetone, methanol and t-butanol.

It is relatively straightforward to evaluate isothermal compressibilities and partial molecular volumes by alternative routes for comparison with the Verlet method results, as described in Section 6.4.2. It is however not a standard procedure to obtain activity coefficient derivatives from MD simulations, and there seems to be no previous study reporting simulation results for these mixtures and using the same force fields. The results obtained in this section are therefore not mainly presented for the sake of validating the Verlet method, but rather to apply the method to determine simulation-based excess Gibbs energy model parameters. Based on the successful validation of the Verlet method carried out in Sections 6.1–6.3, it is here assumed that the method yields accurate results also for molecular mixtures.

In Section 6.4.3, the results obtained by the Verlet method are compared against values derived from correlations of experimental data. This serves however only partially as a validation of the Verlet method, since the comparison depends on the accuracy of the force field as well.

6.4.1 Simulation details

Simulations were carried out of the three mixtures at compositions listed in Table 6.5, using NAMD (Phillips *et al.*, 2005) with the CHARMM27 force field (MacKerell Jr. *et al.*, 1998). For methanol, there were no missing parameters, but for t-butanol, needed values were taken from similar atom types defined in the CHARMM27 force field. For acetone, the CHARMM parameters reported by Martin and Biddy (2005) were used. All parameters are listed in Appendix A. The TIP3P model adapted for CHARMM with flexible bonds was used for water (MacKerell Jr. *et al.*, 1998). The simulations were carried out in the NPT ensemble (molecule number, pressure, and temperature were constant) at a temperature of 323.15 K and a pressure of 1 atm, with each system consisting of 3000 molecules in total. The Langevin thermostat with a damping constant of 5 ps^{-1}, and Langevin piston with a period of 200 ps and a decay constant of 500 ps were employed for controlling respectively temperature and pressure. Periodic boundary conditions were applied in x, y, and z directions, and the particle-mesh Ewald method was employed for calculation of electrostatic forces with a grid point spacing smaller than 1 Å. LJ forces were evaluated using a cutoff distance, switching distance and neighbor list outer radius of respectively 12, 10 and 14 Å. Center-of-mass radial distribution functions were sampled with a bin size of 0.1 Å for values of r up to half the dimension of the simulation box. All simulations were carried out for 9 ns, of which the last 8 ns were used for the property calculations. The dielectric constant was evaluated via the fluctuation in

the total dipole vector (Allen and Tildesley, 1987) and statistical uncertainties were estimated using the blocking method assuming blocks of 2 ns to be independent (Allen and Tildesley, 1987).

Table 6.5: Summary of the binary mixtures simulations. The compositions are given by the fraction of water molecules x_1. Also given are the matching distances R_{ij} used with the Verlet method.

System	x_1	R_{11} [Å]	R_{12} [Å]	R_{22} [Å]
water/acetone	0.10	13.1	14.2	15.0
	0.20	13.2	14.3	15.3
	0.35	13.4	14.4	12.7
	0.50	13.3	15.1	12.3
	0.65	13.1	10.5	13.0
	0.80	13.5	13.0	13.0
	0.90	12.1	11.5	13.6
water/methanol	0.10	12.0	12.8	13.5
	0.20	10.9	11.3	11.8
	0.35	11.8	12.9	13.1
	0.50	12.0	13.3	13.1
	0.65	11.1	13.9	14.1
	0.80	11.6	10.4	9.2
	0.90	10.4	9.2	9.8
water/t-butanol	0.10	12.2	13.3	13.3
	0.20	12.9	13.4	13.5
	0.35	13.3	13.8	13.7
	0.50	13.6	14.1	14.3
	0.65	14.0	14.3	14.2
	0.80	14.8	14.9	15.3
	0.90	15.0	16.3	16.7

6.4.2 Self-Consistency and Comparison with Previous Integration Methods

The RDFs obtained from the simulations of the three mixtures were extended by the Verlet method and integrated numerically. Equations (4.10)–(4.12) were employed to evaluate respectively the isothermal compressibilities, partial molecular volumes and activity coefficient derivatives.

As with the analysis of the LJ/Stockmayer mixtures, the previous integration methods simple truncation (Weerasinghe and Smith, 2003) and the Hess method (Hess and van der Vegt, 2009) were employed to evaluate the same properties. Simple truncation was here employed averaging the integral with R_{lim} varied in the intervals 10–14.5 Å, 9–12.5 Å and 10–15 Å for respectively water/acetone, water/methanol and water/t-butanol. With the Hess method, the scaling factors α_{ij} were evaluated from the calculated RDFs with the parameter R set to 18 Å, 15

Å and 20 Å (see Equation (4.16)) for respectively water/acetone, water/methanol and water/t-butanol. Numerical integration of the rescaled RDFs did still not converge within the sampling range. The integrals of the rescaled TCFs were therefore evaluated by the truncation approach using intervals of 14–18 Å, 13–16.5 Å and 14–19 Å for water/acetone, water/methanol and water/t-butanol, respectively. As in Section 6.2.3, the truncation radii employed for integration of the re-scaled TCFs were larger than those used with the simple truncation approach since the rescaled TCFs probably were more accurate than the original ones for large r.

For comparison, isothermal compressibilities were evaluated via the fluctuations of the simulation box volume (Equation (6.13)) and the results are shown in Figures 6.10(a)–(c). For water/acetone, the Verlet method was in good agreement with the fluctuation formula with discrepancies of at most 7 % (Figure 6.10(a)). Simple truncation and the Hess method did however yield results that overestimated the fluctuation formula values with up to 95 % and 55 %, respectively. For water/methanol, isothermal compressibilities obtained from the fluctuation formula agreed well with those obtained from the Verlet method and simple truncation (Figure 6.10(b)) with discrepancies of less than 7 %. The Hess method yielded also good results at low x_1, but diverged from the fluctuation results for $x_1 > 0.5$. For water/t-butanol, the Verlet method reproduced the fluctuation formula results to within 5 %, while simple truncation and the Hess method failed (Figure 6.10(c)). In fact, the Hess method, as applied here, yielded negative compressibilities. The shortcomings of simple truncation and the Hess method observed for water/acetone and water/t-butanol are probably due to that the RDFs of these mixtures are of longer range than those for water/methanol. Therefore, one probably has to consider simulations of larger systems in order to reliable results by these methods.

In order to validate the partial molecular volumes obtained by the different integration methods, the excess molecular volume, v^{E}, was evaluated for the simulations at each composition according to

$$v^{\mathrm{E}}(x_1) = v(x_1) - x_1 v_1 - x_2 v_2 \tag{6.14}$$

where $v(x_1)$ denotes the average molecular volume obtained at the composition x_1. x_1 and x_2 denote the fractions of respectively water and organic solvent molecules and v_1 and v_2 denote the average molecular volumes of the corresponding pure components, and were obtained from separate simulations. For each of the three mixtures studied, the polynomial model

$$\left(v^{\mathrm{E}}\right)_{\mathrm{model}} = x_1 x_2 \left(a_0 + a_1(x_2 - x_1) + a_2(x_2 - x_1)^2\right) \tag{6.15}$$

was fitted to calculated values of v^{E}. This model was advocated by Handa and Benson (1979). Reduced partial molecular volumes were evaluated by analytical differentiation of the model according to

$$\rho\bar{v}_1 = \rho v_1 + \rho \left(\frac{\partial(N v^{\mathrm{E}})}{\partial N_1}\right)_{T,P,N_2} \tag{6.16}$$

where ρ and N denote respectively the number density and the total number of molecules and N_1 and N_2 denote the number of molecules of component 1 and

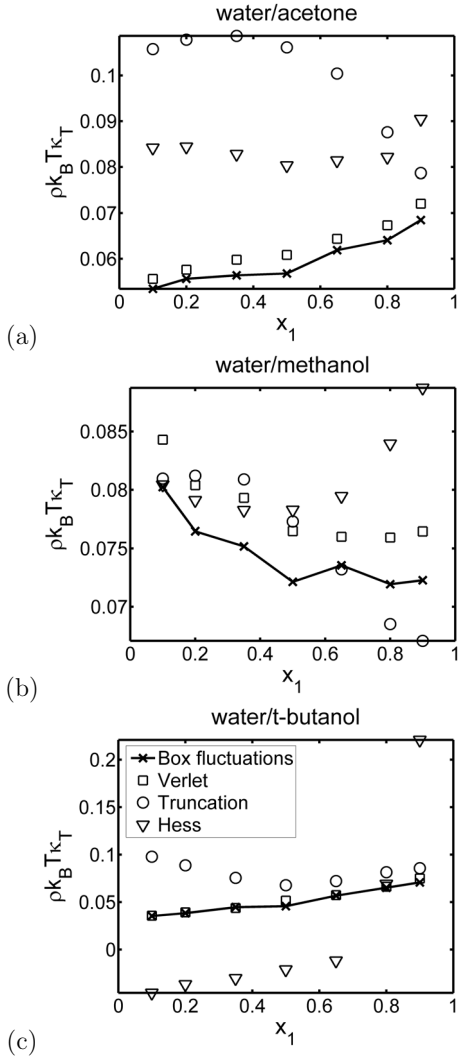


Figure 6.10: Isothermal compressibilities for the mixtures (a) water/acetone, (b) water/methanol (c) and water/t-butanol. The results from the Verlet method (squares), simple truncation (circles) and the Hess method (triangles) for obtaining the TCFIs are compared with values obtained from the fluctuation formula (Equation (6.13)) (crosses + line).

2, respectively. In Figures 6.11(a)–(c), the results are compared with the results obtained via the TCFIs calculated by the Verlet method, simple truncation and the Hess method. Also shown in these figures are values derived from correlations of experimental data which are described in Section 6.4.3.

For all three mixtures, the partial molecular volumes obtained from the correlations of simulation volumes were in very good agreement with those obtained from experimental correlations. The results obtained via the three TCFI calculation methods agreed very well with both correlations. The three methods furthermore yielded similar results. There was however a particular tendency for simple truncation to diverge from the other methods and from experimental data at small x_1. For water/acetone, this resulted in an overestimated $\rho\bar{v}_1$ (Figure 6.11(a)), while for water/t-butanol, simple truncation yielded underestimates (Figure 6.11(c)).

6.4.3 Comparison with Experimental Correlations

Neither TCFIs of DCFIs can be measured directly in experiments, but have to be derived from correlations of experimental data for other thermodynamic properties. A procedure for this was described by Wooley and O'Connell (1991), in which one extracts the isothermal compressibility, partial molecular volumes and activity coefficient derivatives, i.e. the target properties of Equations (4.10)-(4.12), from experimental data. The activity coefficient derivatives are obtained by fitting mixture vapor-liquid equilibrium data to obtain parameters for at least two different G^{E} models. Wooley and O'Connell (1991) employed the Wilson, non-random two liquid (NRTL) and modified Margules (mM) models for this task. Partial molecular volumes are obtained from correlations of mixture densities. Isothermal compressibilities are either taken from measurements or calculated from the accurate correlation of Huang and O'Connell (1987), which is fitted to screened experimental data. The properties are converted into DCFIs, according to

$$C_{11} = 1 - \frac{\rho\bar{v}_1^2}{\kappa_T RT} - x_2 \left(\frac{\partial \ln \gamma_1}{\partial x_1} \right)_{T,P,N_2} \tag{6.17}$$

$$C_{12} = 1 - \frac{\rho\bar{v}_1\bar{v}_2}{\kappa_T RT} + x_1 \left(\frac{\partial \ln \gamma_1}{\partial x_1} \right)_{T,P,N_2} \tag{6.18}$$

$$C_{22} = 1 - \frac{\rho\bar{v}_2^2}{\kappa_T RT} - x_1 \left(\frac{\partial \ln \gamma_2}{\partial x_2} \right)_{T,P,N_1} \tag{6.19}$$

From the DCFIs, the TCFIs can be evaluated by solving the integrated OZ equation (O'Connell, 1971b)

$$(\mathbf{I} + \mathbf{XH})\,(\mathbf{I} - \mathbf{XC}) = \mathbf{I} \tag{6.20}$$

which is a regular systems of linear equations in the TCFIs, in which $\mathbf{I}$ denotes the identity matrix.

For water/acetone, Wilson and NRTL parameters were obtained by regression the Pxy measurements at 45 °C by Taylor (1900). The same treatment was employed for water/methanol, analyzing Pxy measurements at 50 °C by Kurihara *et al.* (1995). These data are of high quality (2 on the van Ness (1995) scale which is based on thermodynamic consistency). For water/t-butanol, Px measurements at 50 °C by

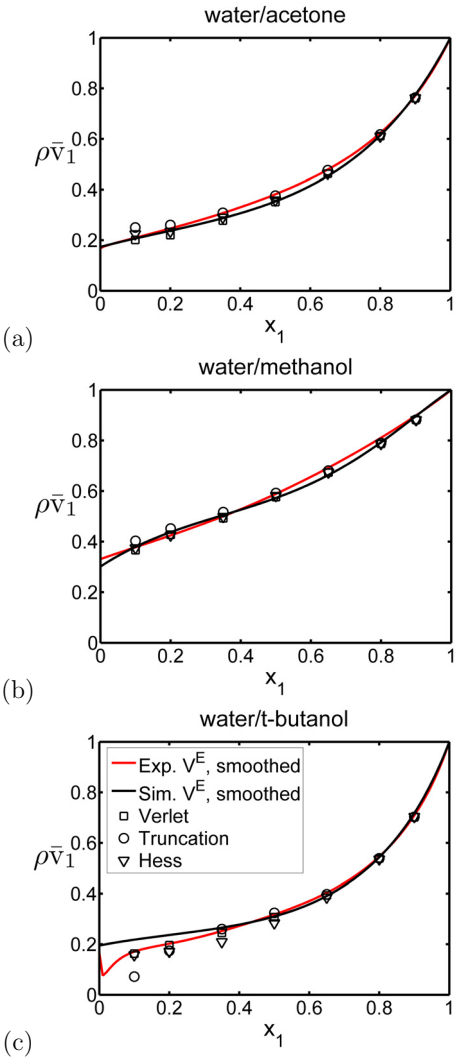


Figure 6.11: Partial molecular volumes for the mixtures (a) water/acetone, (b) water/methanol and (c) water/t-butanol. The results from the Verlet method (squares), simple truncation (circles) and the Hess method (triangles) for obtaining the TCFIs are compared with results obtained by smoothing excess volumes from simulation using a quadratic polynomial (Equation (6.15)) (black line) and smoothed experimental data (red line). Note that the experimental data in (a) and (c) are from measurements at respectively 25 and 55 °C.

Fischer and Gmehling (1994) were considered. These data were regressed using the Wilson and NRTL models, as well as with the 4-parameters mM model (van Ness, 1995). The activity coefficient derivatives for water obtained from the regressions are shown in Figures 6.12(a)–(c).

Partial molecular volumes were evaluated using the v^{E} correlation of Mikhail and Kimel (1961) for water/methanol at 50 °C and Handa and Benson (1979) for both water/acetone at 25 °C and water/t-butanol at 55 °C. The results are shown in Figures 6.11(a)–(c).

Isothermal compressibilities were obtained from the correlation by Huang and O'Connell (1987) since no accurate high-pressure PV measurements seems to be available for the studied mixtures. Since no mixture parameters were available for the mixtures, the mixture isothermal compressibility was approximated by the compressibility of the corresponding ideal solution, as advised by the authors. The ideal-solution compressibility $\kappa_{T,\mathrm{IS}}$ is given by

$$\kappa_{T,\mathrm{IS}} = \frac{x_1 v_1 \kappa_{T,1} + x_2 v_2 \kappa_{T,2}}{x_1 v_1 + x_2 v_2} \tag{6.21}$$

where v_1 and v_2 denote the pure-component molecular volumes of component 1 and 2, respectively, and $\kappa_{T,1}$ and $\kappa_{T,2}$ denote the pure-component compressibilities. These compressibilities were evaluated from the EOS using the corresponding pure-component parameters given by Huang (1986). The assumption that κ_T is well approximated by $\kappa_{T,\mathrm{IS}}$ is reasonable if the excess volumes are small. This is true for the present mixtures, as the excess volume accounts for less than 4.4 %, 4.2 % and 1.7 % of the total mixture volume for water/acetone, water/methanol and water/t-butanol, respectively. These numbers were obtained from the v^{E} correlations of Mikhail and Kimel (1961) (water/methanol) and Handa and Benson (1979) (water/acetone, water/t-butanol).

Water/acetone Figure 6.12(a) shows that for water/acetone, the activity coefficient derivatives obtained from the Verlet method overestimated the experimental correlations which resulted in a poor agreement. That the Wilson and NRTL correlations yielded almost identical results shows that the treatment of experimental data was robust and does not account for the disagreement with the simulations.

As stated above, discrepancies between experiments and simulations might be due to an inaccurate force field description of acetone and/or water. As will be shown below, simulations of the mixtures water/methanol and water/t-butanol agreed satisfactorily with experimental correlations. Therefore, it seems at first that the force field for acetone is inaccurate, rather than the force field for water. The non-bonded CHARMM parameters for acetone reported by Martin and Biddy (2005) were not optimized to reproduce thermodynamic properties, but rather taken from similar atom types. In particular, the parameters including partial charges for the C and O atoms of the C=O group were taken from the C=O group of asparagine and glutamine side chains. In these molecules, the sp^2 C atom binds to one CH_2 group and one NH_2 group. This is different from the acetone structure, in which the sp^2 C atom binds to two CH_3 groups. Due to this mismatch, the parameters might not be accurate for describing acetone-water interactions. In particular, the C and O atoms have partial charges of respectively $+0.55e$ and $-0.55e$, where e denotes the

elementary charge. This is quite different from the OPLS All Atom parameters for acetone, which also were evaluated in the study of Martin and Biddy (2005). For this force field, the C and O atoms have partial charges of respectively $+0.47e$ and $-0.47e$. Although partial charges of different force fields cannot be compared directly, the difference seems quite significant. It is therefore possible that better results for water/acetone mixtures would be obtained if the parameterization of acetone would be treated more rigorously.

The individual TCFIs obtained from the Verlet method were in poor agreement with the values obtained by the Wooley/O'Connell analysis of experimental data. This is caused by the same factors causing the poor agreement of activity coefficient derivatives. The individual TCFIs for the water/acetone system are therefore not shown.

Water/methanol The activity coefficient derivatives for water/methanol obtained from the Verlet method simulations were in reasonable agreement with both the Wilson and NRTL regressions with discrepancies within 1-2 standard errors (Figure 6.12(b)). The simulations seemed to produce slight overestimates. However, the results can be considered satisfactory considering that FST modeling is best suited for strongly non-ideal systems rather than nearly ideal ones (Christensen *et al.*, 2007c), like water/methanol.

The individual TCFIs were in reasonable agreement with those extracted from data as shown in (Figure 6.13(a)–(c)). Errors for the water/water TCFI (H_{11}) were seen for $x_1 = 0.35$ and methanol/methanol (H_{22}) at water-rich compositions (Figure 6.13(c)).

Water/t-butanol For water/t-butanol, there were discrepancies among the G^{E} models employed to smooth the experimental data (Figure 6.12(c)). The NRTL and mM models predicted a phase split since they crossed the boundary curve for phase stability (indicated in Figure 6.12(c)). The Wilson model, by virtue of its form, predicted full miscibility, which is consistent with experiments. Thus, it is unclear precisely what the quantitative values should be. The results from the Verlet method were in good agreement with the NRTL and mM results at mid-range compositions while at high water concentrations, they agreed very well with the Wilson model. The value obtained at $x_1 = 0.1$ looks suspicious and is probably due to the problem of dilute solution simulations also observed in the LJ/Stockmayer mixtures (Section 6.2.2).

For water/t-butanol, the TCFIs obtained from the Verlet method are in good agreement with the NRTL and mM regressions when $x_1 \leq 0.5$ and with the Wilson model at higher water concentrations (Figure 6.14(a)–(c)). The TCFIs obtained by the Verlet method did not seem to diverge at any composition, which indicates that the simulations predicted phase stability at all compositions.

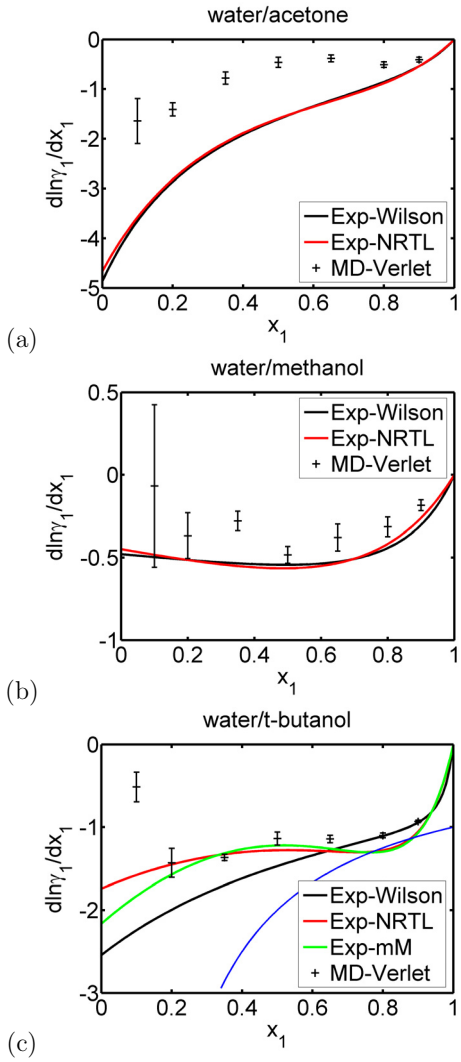


Figure 6.12: Composition derivative of the activity coefficient for the water component vs. the water mole fraction x_1 for the mixtures (a) water/acetone, (b) water/methanol and (c) water/t-butanol. The Verlet method results (crosses) are compared with experimental data smoothed using the Wilson (black line), NRTL (red line) or mM (green line, only (c)) model. In (c), the curve $y = -x_2^{-1}$ is shown (blue line). For full miscibility, activity coefficient derivatives must everywhere lie above this curve. Note also that the experimental data in (a) are from measurements at 45 °C.

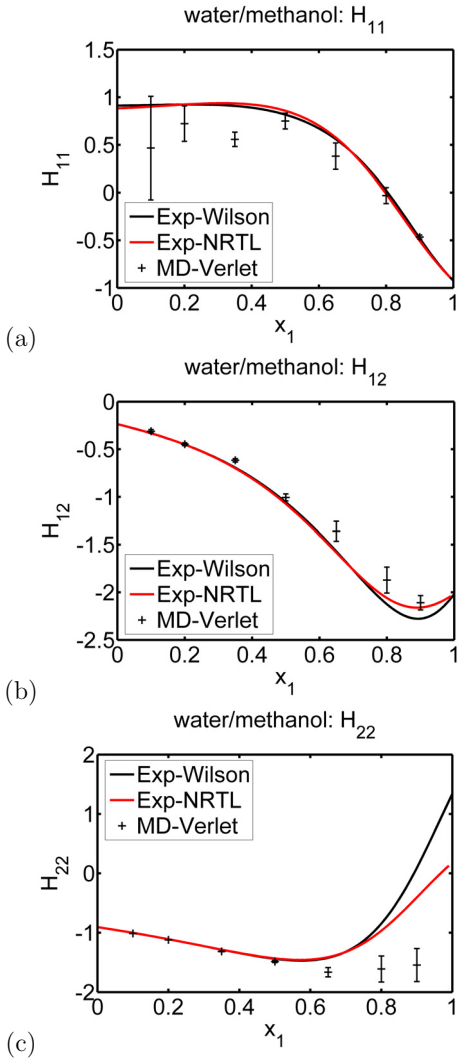


Figure 6.13: TCFIs for (a) water/water, (b) water/methanol and (c) methanol/methanol obtained from simulation of water/methanol mixtures using the Verlet method (crosses) vs. the water mole fraction x_1, compared with TCFIs obtained from experimental data using the procedure of Wooley and O'Connell (1991), where either the Wilson (black line) or NRTL (red line) model was employed for obtaining the activity coefficient derivatives.

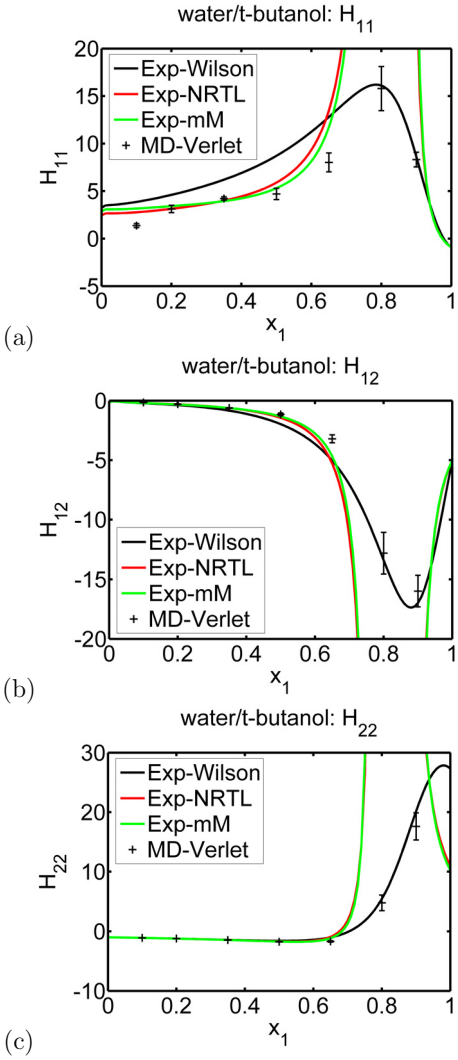


Figure 6.14: TCFIs for (a) water/water, (b) water/t-butanol and (c) t-butanol/t-butanol obtained from simulation of water/t-butanol mixtures using the Verlet method (crosses) vs. the water mole fraction x_1, compared with TCFIs obtained from experimental data using the procedure of Wooley and O'Connell (1991), where either the Wilson (black line), NRTL (red line) or mM (green line) models were employed for obtaining the activity coefficient derivatives. Note that the NRTL and mM model approaches infinity since they predict a phase split.

6.4.4 Regression of Excess Gibbs Energy and Water Activity Calculation

The procedure of Christensen *et al.* (2007c) described in Section 4.2.3 was employed to fit the mM model for the excess Gibbs energy to the activity coefficient derivatives obtained by the Verlet method. The four-parameter mM model (Abbott and van Ness, 1975) was selected for all three mixtures. For water/t-butanol, the activity coefficient derivative obtained at $x_1 = 0.1$ seemed unreliable, as discussed above, and was therefore not included in the curve fitting.

The activity coefficient γ_1 for the water component was via Equation (4.23) evaluated as a function of x_1 for all three mixtures. The corresponding water activities $a_1 \equiv \gamma_1 x_1$ are shown in Figure 6.16. Water/t-butanol is here predicted to be the most non-ideal mixture followed by water/acetone and water/methanol.

6.5 Summary

The previous chapter described a computational methodology for extending RDFs obtained from molecular simulation to arbitrarily large spatial separations, so that TCFIs can be reliably obtained by numerical integration. In this chapter, numerical tests have been carried out in order to validate the accuracy of the calculated TCFIs. The Verlet method was tested for analyzing simulations of pure atomic fluids, binary atomic mixtures, pure molecular fluids and binary molecular mixtures. The computed TCFIs were validated by comparing certain "target" derivative properties with values obtained from alternative approaches, or from EOS fitted to previous simulations. The tests showed that the TCFIs are obtained with a good accuracy. Less good accuracy might be obtained if the system is near the critical point of if a binary mixtures is simulated at a composition where either component is dilute. A third limitation might be encountered as the molecules become less spherical, since the assumptions underlying the methodology may become less valid. Nevertheless, good results were obtained for molecules as asymmetrical as hexane. The treatment of polar molecules such as water and 2-propanol seemed to be accurate as well.

The Verlet method compared favorably with two previously suggested approaches for TCFI calculation, namely simple truncation (Weerasinghe and Smith, 2003) and the Hess method (Hess and van der Vegt, 2009). If the simulated system is sufficiently large, as was the case here for water/methanol, the three methods can be expected to yield similar results. The Verlet method is however superior when the system is small and the RDFs have significant structure beyond the sampling limit imposed by the simulation box dimensions, as presently was the case for the LJ/Stockmayer, water/acetone and water/t-butanol mixtures. This is a significant result since the Verlet method might allow thermodynamic derivative properties to be accurately obtained from simulations of small systems, which can be carried out with relatively low computational efforts.

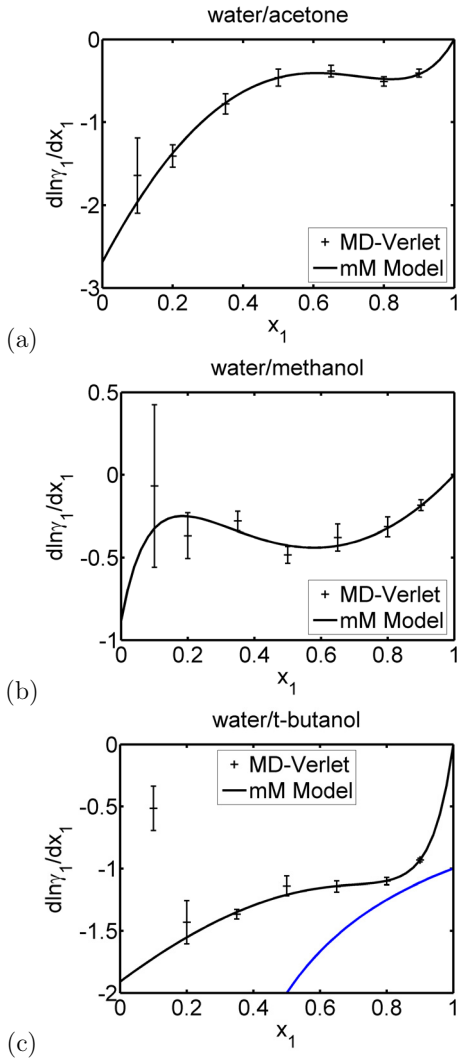


Figure 6.15: Composition derivative of the activity coefficient for the water component vs. the water mole fraction x_1 for the mixtures (a)water/acetone, (b) water/methanol and (c) water/t-butanol, obtained from the four-parameters mM model (line) fitted to the values obtained from the simulations using the Verlet method (crosses). The data point at $x_1 = 0.1$ in (c) was not considered in the regression of the model, as elaborated in the text. In (c), the curve $y = -x_2^{-1}$ is shown (blue line). For full miscibility, activity coefficient derivatives must everywhere lie above this curve.

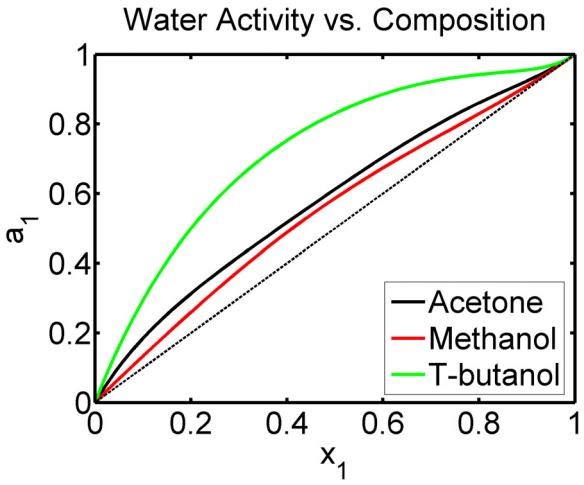


Figure 6.16: Thermodynamic water activity a_1 as a function of water mole fraction x_1 for the mixtures water/acetone (solid black line), water/methanol (red) and water/t-butanol (green). The curves were derived from the mM model fitted to activity coefficient derivatives obtained from the simulations. The line $a_1 = x_1$ is also shown (dotted black line) in order to indicate how much these mixtures deviate from ideal solution behavior.

7

Molecular Dynamics Study of *Candida Antarctica* Lipase B - Part II

The molecular dynamics (MD) study of *Candida antarctica* lipase B (CALB) in water and organic solvents described in Chapter 3 is extended in this chapter. Five organic solvents are considered. Three of them are polar, namely acetone, methanol and t-butanol, while two of them are non-polar, namely methyl t-butyl ether (MTBE) and hexane. The purpose of the investigations is first to demonstrate how the developments of Chapters 4–6 can be applied to determine the bulk water activity in MD simulations of proteins in non-aqueous media. The second aim is to determine how the structure and dynamics of CALB depends on the water activity. The significance of water activity for CALB in a gas/solid reactor was previously addressed by Branco *et al.* (2009). The contribution of this investigation is to study the effect of water activity in the presence of organic solvent. This allows as well studying how different organic solvents directly influence the structure and dynamics of CALB.

The solvents considered in this chapter are except for methanol commonly used in non-aqueous biocatalytic systems with CALB (Anderson *et al.*, 1998). T-butanol (and t-pentanol) is sometimes said to be the preferred solvent since CALB is claimed to be especially stable in this solvent (Wang *et al.*, 2006; Su and Wei, 2008; Fjerbaek *et al.*, 2009). MTBE was in the transesterification study of Abildskov *et al.* (2010a) shown to induce a relatively high activity of CALB which could not be rationalized in terms of the shift of reaction equilibrium (see Figure 2.1). Methanol is not used as a solvent for CALB as it has been observed to inactivate the enzyme (Kaieda *et al.*, 2001; Wang *et al.*, 2006). The mechanism of this inactivation is not understood. It is therefore of interest to investigate how these solvents affect CALB on the molecular scale. CALB has previously been studied in methanol by MD simulation (Trodler and Pleiss, 2008). It seems however that CALB has not previously been studied in the four remaining solvents.

The setup and performance of the MD simulations are described in Section 7.1. The distribution and dynamics of water and organic solvent molecules in the simulations are analyzed in Section 7.2. The effects on CALB structure and flexibility are discussed in Sections 7.3 and 7.4, respectively.

7.1 Simulation Procedure

MD simulations were carried out of CALB in acetone, methanol, t-butanol, MTBE, hexane and pure water. For each of the five organic solvents, simulations were carried out at five different hydration levels in order to study the structure and flexibility of CALB on a broad range of water activities. Section 7.1.1 describes the initial setup of the simulations and the selection of force fields while Section 7.1.2 describes how the MD simulations were carried out.

7.1.1 System Setup and Force Fields

The simulated systems were set up by a procedure similar to the one used in Chapter 3. CALB coordinates were taken from the best resolved crystal structure 1TCA (Uppenberg *et al.*, 1994). Protonation states were chosen as in Section 3.2.2. For the setup of systems containing organic solvent, a water layer was for convenience first built around the protein. For each of the five organic solvents, five systems were prepared with water layers of different sizes. For water layers containing less than 286 molecules (which is the number of crystal waters in 1TCA), the crystal waters with lowest B-factors were retained. For water layers containing more molecules, all crystal waters were retained and additional water molecules were placed around CALB using the VMD plug-in SOLVATE (Humphrey *et al.*, 1996). The CALB/water complex was embedded in a cubic box containing one of the five organic solvents. The organic solvent molecules were taken from the last frame of an MD simulation of the pure organic solvent of at least 500 ps. Organic solvent molecules closer than 2.5 Å to the CALB/water complex were removed. For the systems containing CALB in pure water, the crystal waters were retained and a water box was built using SOLVATE (Humphrey *et al.*, 1996). The simulations are listed in Table 7.1 along with the number of water and organic solvent molecules included in each system.

For the water-miscible solvents acetone, methanol and t-butanol, the number of water molecules in the layer were selected in order to achieve water activities spanning the range from 0 to 1 as good as possible. As discussed in Chapter 4, the number of water molecules required to attain a given activity was not known *a priori*, since the molecules at equilibrium are distributed over the bulk phase and the protein surroundings. The selection was therefore guided by experience from previous CALB simulation and activity calculations using UNIFAC (Hansen *et al.*, 1991). As will be shown in Section 7.2.1, the prepared systems turned out to span the activity range quite well for all three organic solvents.

As apparent from Table 7.1, the systems containing water-miscible organic solvent were designed to contain around 40000 atoms, which was more than enough to ensure that the protein molecule did not interact with any of its periodic images. These system sizes were used in order to ensure that accurate estimates of the fraction of water in the bulk region could be obtained. Previous experience e.g. from the acetone systems of Chapter 3 suggested that accurate estimates could not be obtained from simulation of smaller systems. For the systems containing MTBE or hexane, the bulk water activity was not evaluated. It was therefore not necessary to design those systems as large as the water-miscible solvent systems.

The CHARMM27 force field (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004)

was used for all simulations. Protein, water, acetone and hexane molecules were modeled as described in Section 3.2.2. Methanol and t-butanol were modeled by the CHARMM27 force field parameters, while the CHARMM35 parameters (Vorobyov *et al.*, 2007) for ethers were employed for MTBE. For MTBE and t-butanol, there were missing parameters. For these molecules, parameters for similar atom types were used. All organic solvent parameters are listed in Appendix A.

7.1.2 Simulation Details

The simulations were carried out using the simulation program NAMD (Phillips *et al.*, 2005). The systems were minimized with respect to the total configurational energy and simulated using the same simulation parameters as in Section 3.2.3, with a few exceptions. The temperature was here set to 323.15 K (50°C) since 298.15 K (25°C) is roughly the temperature where t-butanol crystallizes. Each simulation was carried out for a total of 20 ns and the last 10 ns were used for the analysis. The longer simulation time was used since the number of water molecules in the first solvation shell of CALB required up to 10 ns of simulation to equilibrate in the t-butanol systems, as will be shown in Section 7.2. Due to the slow equilibration, the initial period where the simulation was run with the C_α atoms constrained was prolonged as well. During the first nanosecond, the C_α atoms were constrained to their initial positions. During the following nanosecond, the same atoms were restrained by a harmonic potential with a force constant of 1 kcal/mol/Å^2. This was followed by 18 ns of unconstrained simulation.

Three replica simulations which were started from different initial velocities were carried out of each system in order to estimate statistical uncertainties in the results. For the pure water system, five replica simulations were carried out.

7.2 Hydration and Solvation

The hydration of CALB was first monitored by counting the number of water molecules in the first solvation shell, as defined in Section 3.3. For the simulations carried out in acetone, methanol or t-butanol, this number decreased initially and stabilized at an approximately constant level, as the molecules of the water shell around CALB mixed with the bulk solvent. This is similar to the behavior seen in Figure 3.4(a)–(b). The constant level was reached after approximately 4, 5 and 10 ns for simulations carried out in methanol, acetone and t-butanol, respectively. These times did not depend significantly on the total number of water molecules in the system. In hexane and MTBE, the hydration level remained constant throughout the simulation, since essentially all water remained near the protein surface. This behavior was previously shown in Figure 3.4(c). In a few hexane simulations, and in most MTBE simulations, one or two water molecules were observed to escape from the hydration layer around the protein and diffuse through the bulk medium. For the pure water simulations, the hydration level increased slightly in the beginning, but reached a constant level after 4 ns, which is consistent with the previous simulations (Figure 3.4(d)).

Table 7.1: Summary of CALB simulations listing number of water and organic solvent molecules, total number of atoms and a short identifier for each simulation. For all systems, three replica simulations which were started from different initial velocities were carried out. For the pure water system, five such simulations were carried out.

Solvent	#water	#solvent	#atoms	ID
Acetone	135	3700	42030	A135(a)-(c)
	335	3640	42030	A335(a)-(c)
	920	3460	41985	A920(a)-(c)
	2300	3050	42025	A2300(a)-(c)
	5900	1970	42025	A5900(a)-(c)
Hexane	43	1250	29775	H43(a)-(c)
	87	1250	29886	H87(a)-(c)
	171	1250	30138	H171(a)-(c)
	335	1250	30630	H335(a)-(c)
	500	1250	31125	H500(a)-(c)
Methanol	210	6125	42005	M210(a)-(c)
	940	5760	42005	M940(a)-(c)
	2500	4980	42005	M2500(a)-(c)
	4200	4130	42005	M4200(a)-(c)
	6135	3160	41990	M6135(a)-(c)
MTBE	50	1350	29075	E50(a)-(c)
	100	1350	29225	E100(a)-(c)
	171	1350	29438	E171(a)-(c)
	335	1350	29930	E335(a)-(c)
	500	1350	30425	E500(a)-(c)
T-butanol	65	2480	42020	T65(a)-(c)
	130	2470	42065	T130(a)-(c)
	210	2450	42005	T210(a)-(c)
	700	2350	41975	T700(a)-(c)
	2970	1870	41585	T2970(a)-(c)
Water	9500	-	33125	W(a)-(e)

7.2.1 Bulk Solvent Composition and Water Activity

In all simulations, the water molecules were to some extent partitioned between the hydration layer of the enzyme and the bulk medium. In order to quantify this behavior, the water content of the bulk medium was determined. This is as well a necessary step of the *A posteriori* approach to determine the water activity of the system, as described in Section 4.2. It was for the calculation required that the simulation box as shown in Figure 7.1 was divided into two regions, the "bulk" region and the "protein vicinity" region, in the following referred to as region I and II, respectively. A water or organic solvent molecule was defined to be in region I, if its distance to the protein surface was greater than a selected boundary distance R_{bound}. Likewise, the molecule was by definition in region II, if its distance to the protein surface was smaller than R_{bound}. For a water molecule, the distance to the protein surface was defined as the distance from the O atom to the closest non-hydrogen atom of the protein. For the organic solvent molecules, the distance was measured from a "central" C atom, i.e. the carbonyl C atom of acetone, the tertiary C atom of MTBE and t-butanol, and the (only) C atom of methanol. For hexane, one of the two C atoms in the middle of the chain was chosen and employed consistently. The boundary distance R_{bound} needs to be chosen sufficiently large

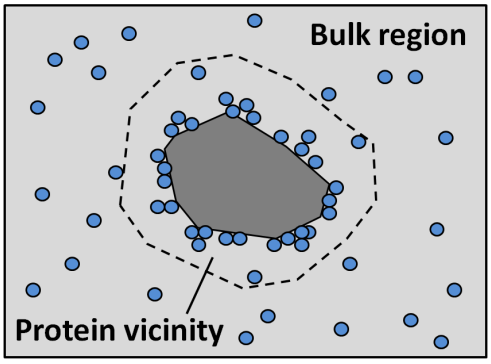


Figure 7.1: Illustration of how the simulation box is divided into "bulk" (I) and "protein vicinity" (II) regions.

that the water/organic mixture in region I is homogeneous, and thus approximately unaffected by the presence of the protein. In order to properly select this distance, the fraction of water molecules x_{w} was for a set of representative simulations evaluated as a function of the distance r to the protein surface. This was accomplished by counting the average number of water and organic solvent molecules in a shell of thickness Δr at a distance r from the protein surface, denoted by $N_{\text{w},\Delta r}(r)$ and $N_{\text{s},\Delta r}(r)$, respectively. This lead to

$$x_{\text{w}}(r) = \frac{N_{\text{w},\Delta r}(r)}{N_{\text{w},\Delta r}(r) + N_{\text{s},\Delta r}(r)} \tag{7.1}$$

In all systems, $x_w(r)$ varied close to the protein surface but reached a plateau for large r, as shown for selected systems in Figures 7.2(a)–(c). The distance from the protein surface to the homogeneous region was in general different for the different systems, as it increased with the number of water molecules in the system. The plateau was however in all systems reached when $r \geq 10$ Å. Thus, $R_{\text{bound}} = 10$ Å was used consistently to define the boundary between region I and II. The average fraction of water molecules in region I was evaluated, and for the acetone, methanol and t-butanol simulations, the corresponding water activity was determined using the Gibbs energy models regressed in Section 6.4.4.

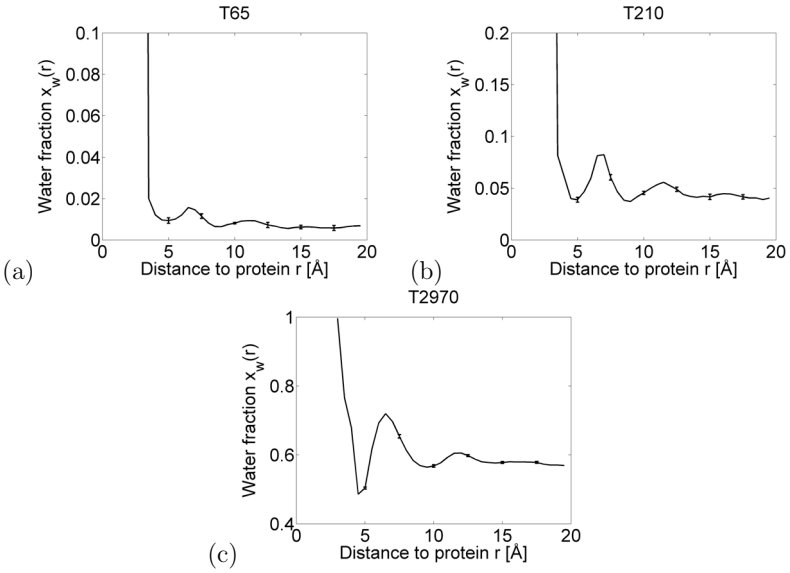


Figure 7.2: The function $x_w(r)$ evaluated using $\Delta r = 0.5$ Å for the three t-butanol systems T65 (a), T210 (b) and T2970 (c). The curves are averages over the three replica simulations for each system. Standard errors (which are quite small) are shown for selected values of r. Note that the scales on the y-axes are different.

7.2.2 Adsorption Isotherms

Water adsorption isotherms for CALB determined from simulations in acetone, methanol and t-butanol are shown in Figures 7.3(a)–(b). The average number of water molecules in the first solvation shell is used to quantify the amount of adsorbed water. If the amount of adsorbed water is plotted against the bulk water fraction x_w as in Figure 7.3(a), there are clear differences between the different solvents. At similar x_w, most water is adsorbed in t-butanol, followed by acetone and methanol. The simulations of the corresponding binary mixtures analyzed in Section 6.4 showed that for the force field parameters used, water mixes most favorably with methanol, followed by acetone and t-butanol. This is consistent with the previous finding that the more polar the solvent is (and thus more water-miscible), the higher the water

content in the solvent needs to be in order to achieve a certain protein hydration level. This is in good agreement with simulations by Yang *et al.* (2004); Micaêlo and Soares (2007) and Trodler and Pleiss (2008).

If the amount of adsorbed water instead is shown in terms of the bulk water activity a_w as in Figure 7.3(b), the adsorption isotherm show less variation with the organic solvent. At low a_w, CALB is slightly less hydrated in methanol than in acetone and t-butanol. At high a_w, CALB is on the other hand less hydrated in t-butanol than in acetone and methanol. These differences can possibly be attributed to that water and organic solvent molecules compete for binding to the protein surface. At low a_w, the adsorption of water is driven by water molecules binding to hydrophilic sites at the protein surface (Soares *et al.*, 2003; Micaêlo and Soares, 2007; Branco *et al.*, 2009) (see also Section 7.2.3). Methanol can probably mimic water better than acetone and t-butanol, due to its smaller non-polar group. The hydration level might thus be lower in methanol as molecules of this type probably are more competitive for binding to the hydrophilic sites. At high a_w, the hydrophilic sites are all, occupied and the adsorption is instead driven by water molecules binding to water cluster already present on the protein surface (Soares *et al.*, 2003; Micaêlo and Soares, 2007; Branco *et al.*, 2009) (see also Section 7.2.3). In this region, the coverage of the protein surface with water molecules might be driven by the removal of organic solvent molecules adsorbed to the hydrophobic surface. T-butanol molecules might bind more tightly to this surface than acetone and methanol, due to their more bulky, non-polar portion. This might explain the lower hydration of CALB observed in t-butanol at high a_w. In order to support this argument, Figure 7.4 shows the average number of water molecules which are located close to hydrophobic residues, but not to hydrophilic residues (as defined in Section 3.3). At similar a_w, this number is apparently highest in methanol and lowest in t-butanol (at low a_w, the data for t-butanol is rather uncertain). This suggests that t-butanol molecules are more difficult to displace from the hydrophobic surface than acetone molecules, which in turn are more difficult to displace than methanol molecules.

Water adsorption isotherms for CALB determined from simulations in hexane and MTBE are shown in Figure 7.5. Although very few water molecules in these simulations are found in the bulk medium, the water fraction can be estimated. The spread in the values obtained from different replica simulations indicates however that the estimates are somewhat uncertain. At low water contents, longer simulations are probably required to obtain accurate estimates. It is nevertheless clear that at similar CALB hydration levels, x_w is significantly higher in MTBE than in hexane. This is reasonable as the solubility of water is higher in MTBE than in hexane.

Despite the small differences between the water adsorption isotherms of acetone, methanol and t-butanol, the conclusion seems to be that in these solvents, approximately similar hydration levels are obtained if the water activities are similar. This is consistent with water adsorption isotherms observed in experiments (Halling, 1990a) and validates the common assumption the hydration level can be controlled in experiments by controlling the water activity (Halling, 1994). One could expect that the corresponding adsorption isotherms in MTBE and hexane would look similar, if expressed in terms of the water activity. This can however not be concluded here, since the methodology used to evaluate the water activities only is applicable to

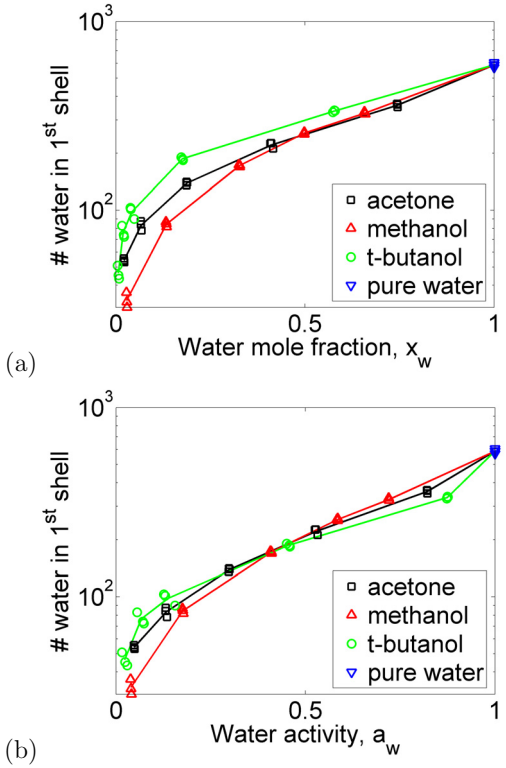


Figure 7.3: Average number of water molecules in the first shell of CALB vs. bulk water fraction (a) and bulk water activity (b). Results are shown for the acetone (black squares + line), methanol (red triangles + line) and t-butanol (green circles + line) simulations. Results from the pure water simulations are used to extrapolate to $x_{\mathrm{w}} = 1$ and $a_{\mathrm{w}} = 1$.The symbols show the results for each individual simulation to indicate the spread, while the lines connect the average values for each system.

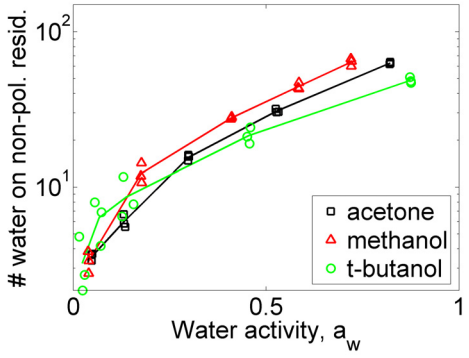


Figure 7.4: Average number of water molecules in the first shell of CALB which are located close to hydrophobic residues, but not close to any hydrophilic residue (as defined in Section 3.3). Colors and symbols are used like in Figure 7.3
.

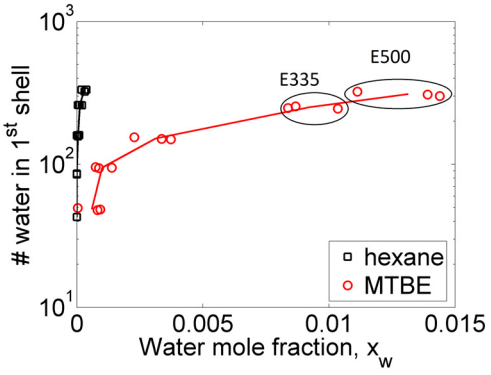


Figure 7.5: Average number of water molecules in the first shell of CALB vs. bulk water fraction. Results are shown for the hexane (black squares + line) and MTBE (red circles + line) simulations. The symbols show the results for each individual simulation to indicate the spread, while the lines connect the average values for each system.

solvents that are miscible with water. It is thus an important part of future investigation to extend the methodology to water-immiscible solvents so that adsorption isotherms such as those shown in Figure 7.3(b) can be studied also in MTBE and hexane.

In the remaining parts of this chapter, results obtained in the different solvents will be reported as functions of the hydration level, so that MTBE and hexane can be compared on the same footing. Corresponding water activities for acetone, methanol and t-butanol can be estimated from Figure 7.3.

7.2.3 Water Clusters at Surface

In experimental measurements of water adsorption isotherms for proteins in organic media, a steep increase of the amount of adsorbed water is often seen at high a_{w}. This may be due to adsorption of water molecules being cooperative as water molecules bind to already formed water clusters on the protein surface. For the present simulations, water clusters were identified by joining water molecules whose O–O distance was less than 3.5 Å. The clusters originating from water molecules in the first solvation shell of CALB were identified. Note that with this definition, the clusters are required to contain at least one water molecule that is in contact with the protein, but may as well contain water molecules that are outside the first solvation shell. The average number of such clusters, and their average size as determined from the simulations is shown in Figures 7.6(a)–(b). The number of clusters at the surface showed a bell-shaped dependence on the hydration level, with a maximum attained at around 150–200. For acetone, methanol and t-butanol, this corresponds to a water activity of 0.4–0.5. The bell shape shows that at low a_{w}, individual water molecules or small water clusters bind to specific sites on the protein surface. As a_{w} increases, the number of clusters at the surface increases until the clusters start to percolate at an a_{w} of 0.4–0.5. This is similar to adsorption isotherms for CALB obtained in the gas phase (Branco *et al.*, 2009). This also resembles the results of Micaêlo and Soares (2007), although they did not observe the number of clusters to decline at high hydration levels. The reason might be that their study did not cover as high hydration levels as here, or that they employed a smaller cutoff distance for joining water molecules to clusters. While all five organic solvents yielded bell-shaped curves attaining the maximum at roughly the same hydration level, the maximum number of clusters was different. This number was ~ 60 for hexane and MTBE, ~ 80 for acetone and t-butanol and ~ 100 for methanol. The surface water was apparently organized in fewer and larger clusters in the non-polar solvents.

The average cluster size seems to be independent of the solvent when the hydration level $\lesssim$ 150–200 ($a_{\mathrm{w}} \lesssim$ 0.4–0.5). At higher hydration levels, the cluster size correlates with solvent polarity since more water is contained in polar solvents.

7.2.4 Water and Organic Solvent Residence Times

The previous sections were focused on static properties of the hydration layer and their dependence on the water activity. In the literature, it has been proposed that also the dynamic properties of the hydration layer are of importance for the protein dynamics. Chapter 3 discussed the hypothesis of Trodler and Pleiss (2008) that

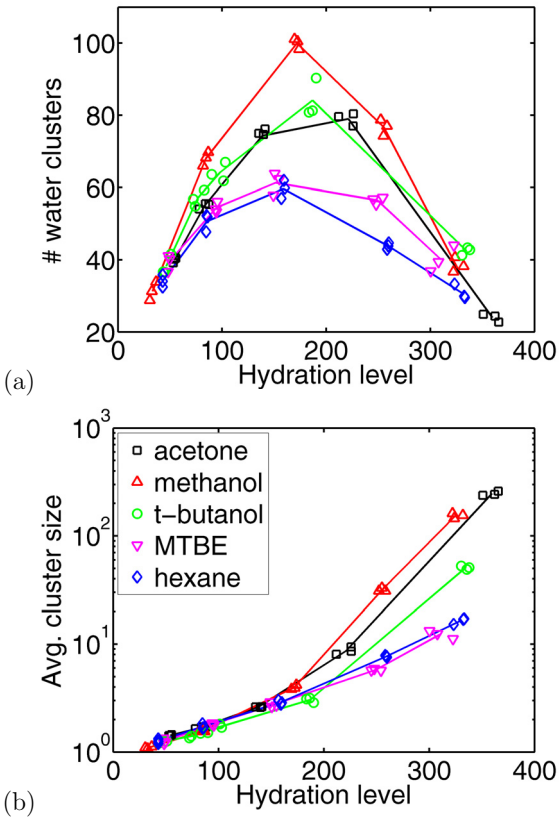


Figure 7.6: (a) Average number of water clusters originating from the surface of CALB and (b) the average size of these clusters vs. the number of water molecules in the first solvation shell. Results are shown for the acetone (black squares + line), methanol (red triangles + line), t-butanol (green circles + line), MTBE (magenta triangles + line) and hexane (blue diamonds + line) simulations. The symbols show the results for each individual simulation to indicate the spread, while the lines connect the average values for each system.

protein flexibility correlates negatively with the number of slowly exchanged water molecules at the surface.. For the present simulations, a more detailed analysis of the water dynamics was carried out by computing residence times. For a water molecule located near a specific protein residue, the residence time is interpreted as the average time elapsed before the water molecule leaves the vicinity of this residue. The formal definition of residence time used here follows that of Makarov *et al.* (2000) and Schröder *et al.* (2006). One introduces first the function $\chi_{i,\alpha}(t)$ for which i and α denote a particular water molecule and protein residue, respectively. The function is defined by

$$\chi_{i,\alpha}(t) = \begin{cases} 1 & \text{if molecule } i \text{ is in the first shell of residue } \alpha \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

As before, a water molecule is defined to be in the first shell if its O atom is within 3.5 Å of any non-hydrogen atom of the protein residue. The function $\chi_{i,\alpha}(t)$ realizes thus a stochastic process for each protein residue and each water molecule that assumes only the values 0 and 1. The autocorrelation function for this process corresponding to the residue α is defined by

$$\rho_\alpha(t) = \frac{1}{T_2 - t - T_1} \int_{T_1}^{T_2 - t} dt_0 \left\langle \chi_{i,\alpha}(t_0 + t)\chi_{i,\alpha}(t_0) \right\rangle_i \tag{7.3}$$

where T_1 and T_2 respectively denote the start and end of the simulation time block used in the calculation, and $\langle\cdot\rangle_i$ denotes averaging over all water molecules in the system, which is employed since all water molecules are equivalent. $\rho_\alpha(t)$ was evaluated from each simulation for t in the interval 0–2.5 ns and for all sufficiently solvent-exposed residues α. Residues with an average exposed surface area fraction of at least 0.25 (see Section 7.3.3 for the definition) were considered to be sufficiently solvent-exposed. The approach allows studying residence times of individual residues but here, the overall autocorrelation function $\rho(t)$ for bound water molecules was calculated by averaging $\rho_\alpha(t)$ over all solvent-exposed residues α. The bi-exponential model of Makarov *et al.* (2000) was then fitted to the function $\rho(t)$. The model is defined by

$$\rho_{\mathrm{bi-exp}}(t) = a_1 e^{-k_1 t} + a_2 e^{-k_2 t} \tag{7.4}$$

where a_1, a_2, k_1 and k_2 are adjustable parameters. $\tau_1 \equiv k_1^{-1}$ and $\tau_2 \equiv k_2^{-1}$ are the residence times of respectively slowly and rapidly exchanged water molecules (thus $\tau_1 > \tau_2$, consistently). The relative populations of slowly and rapidly exchanged water molecules, P_{slow} and P_{rapid}, can be estimated by (Makarov *et al.*, 2000)

$$P_{\mathrm{slow}} = 1 - P_{\mathrm{rapid}} = \frac{a_1}{a_1 + a_2} \tag{7.5}$$

In the pure water simulations, the values $\tau_1 = 0.6 \pm 0.2$ ns, $\tau_2 = 40 \pm 5$ ps and $P_{\mathrm{slow}} = 14.1 \pm 0.8\%$ were obtained. As apparent from Figures 7.7(a)–(b), the residence times of water molecules were longer in the organic solvent simulations. Regardless of the solvent, τ_1 decreased with increasing CALB hydration level. τ_2 followed the same trend, except at low hydration levels where the residence times obtained in MTBE

and hexane seemed to increase with increasing hydration. At high hydration, τ_1 and τ_2 seemed to approach the values obtained in pure water. The lowest values of both τ_1 and τ_2 were obtained in methanol followed by acetone. For hydration levels larger than 100, the t-butanol, MTBE and hexane simulations yielded nearly identical residence times. At lower hydration levels, both τ_1 and τ_2 seemed to be larger in t-butanol than in MTBE and hexane, although statistical uncertainties might account for the differences in τ_1. Since water molecules interact more favorably with polar solvents like methanol and acetone, residence times are expected to be shorter in those solvents than in non-polar ones. This trend was also confirmed by the simulations of cutinase of Micaêlo and Soares (2007). That t-butanol yields residence times that are at least as long as those obtained in MTBE and hexane is more surprising.

The fraction P_{slow} of slowly exchanged water molecules did as well decrease with increasing hydration, regardless of the solvent (Figure 7.7(c). The lowest values of P_{slow} were obtained in methanol followed by acetone. Values obtained in t-butanol, MTBE and hexane were higher and nearly identical. P_{slow} seemed in methanol to approach 40% at low hydration while it approached a value of 75–90% in the other solvents. It is reasonable that higher percentages of loosely bound water are observed in methanol and acetone, since these solvents contain more water than the non-polar solvents when the hydration levels of CALB are similar, as shown in Section 7.2.2. It is however again surprising that t-butanol yields results similar to those obtained in MTBE and hexane. The results demonstrate that the dynamical properties of the water layer are not only a function of the water activity, but also of the organic solvent species.

The residence times for organic solvent molecules were evaluated using the same procedure as for the water molecules. An organic solvent molecule was considered to be bound to the protein if the "central" C atom was within 6.5 Å from any non-hydrogen CALB atom. The "central" C atoms were for the different solvents defined as in Section 7.2.1. The 6.5 Å cutoff was used since radial distribution functions (RDFs) for acetone, t-butanol, MTBE and hexane around protein residues had their first minima at approximately this distance. The same cutoff was for simplicity employed for methanol although the first minima of the corresponding RDFs were at a slightly smaller distance. The bi-exponential model of Equation (7.4) was fitted to the calculated autocorrelation function averaged over all surface-exposed residues, and the residence times and fraction of slowly exchanged molecules was evaluated as above.

Figures 7.8(a)–(b) demonstrate that for acetone, methanol, MTBE and hexane molecules, the residence times were rather insensitive to the hydration level. τ_1 was around 1–2 ns for acetone, methanol and MTBE and τ_2 was one order of magnitude lower. The fraction of slowly exchanged molecules P_{slow} was around 35, 25 and 38 % for acetone, methanol and MTBE molecules, respectively, and approximately independent of the hydration level (Figure 7.8(c)). The residence times of hexane molecules were consistently shorter and also rather insensitive to hydration. The shorter residence times were probably due to that hexane lacks a polar group, and therefore cannot bind tightly to the polar portion of the protein surface. For t-butanol molecules, both τ_1 and τ_2 decreased with increasing hydration (Figure 7.8(a)–(b)). At low hydration, τ_1 and τ_2 were 5.7 ns and 0.36 ns, respectively, which

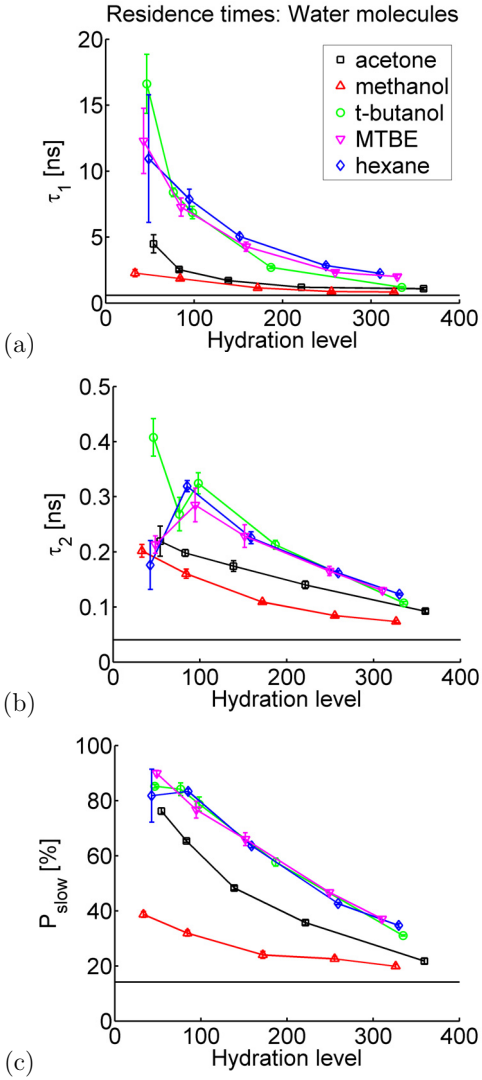


Figure 7.7: Residence times of (a) slowly and (b) rapidly exchanged water molecules vs. the hydration level determined in simulations with acetone (black squares + line), methanol (red triangles + line), t-butanol (green circles + line), MTBE (magenta triangles + line) or hexane (blue diamonds + line) as main solvent. In (c), the corresponding populations of slowly exchanged water are shown. Standard errors estimates were based on three replica simulations which were started from different initial velocities. Corresponding parameters obtained from the pure water simulations are in (a)–(c) shown as black horizontal lines.

is significantly higher than the values obtained in the other solvents. For t-butanol, P_{slow} was 79% at the lowest hydration level and decreased with increasing hydration.

The long residence times observed in t-butanol can possibly be explained if it is assumed that t-butanol molecules via the OH group. The three non-polar methyl groups of the solvent could then act as an "umbrella" shielding the binding site from water molecules and OH groups of other t-butanol molecules that might facilitate the breaking of the bond to the protein. This effect would not be seen in MTBE or hexane as these molecules lack strongly polar groups. It would neither be seen in acetone or methanol since the non-polar portions of these molecules are too small to act as "umbrellas".

The results demonstrate that direct interactions between protein and organic solvent molecules are different for different solvents. These interactions may be modulated by the hydration level, as here was the case for t-butanol. Possible implications of this for the protein flexibility are discussed in Section 7.4

7.3 Structure

7.3.1 Root Mean Square Deviation

The root mean square deviation (RMSD) of CALB with respect to the crystal structure coordinates (1TCA) was monitored for each simulation similarly to Section 3.4.1. Only the C_α atoms were considered in the RMSD evaluation. Since the N- and C terminals (residues 1–20 and 308–317, respectively) caused a drift in the total RMSD in several of the simulations, these regions were consistently omitted from all calculations. In roughly half the simulations, stable RMSD curves were obtained. For the other simulations, stable curves were obtained if one or several flexible regions were omitted from the calculation. Depending on the simulation, these regions were the loop L1 (residues 23–32), the loop L4 (residues 67–75), the helix α5 and adjacent loop segment (residues 138–152) and the loop L13 together with the adjacent helix α10 (residues 243–292). This analysis was carried out for all simulations, and stable RMSD curves were obtained in each case. Selected RMSD plots are shown in Appendix F.

Figures 7.9(a)–(e) show the total RMSD averaged over the last 10 ns of the simulations carried out in the five organic solvents. In the three polar solvents and as well in MTBE, the RMSD was generally lower than in pure water. In t-butanol, the RMSD increased with increasing hydration. The lowest RMSD values (0.99 ± 0.02 Å) were observed in this solvent at low hydration. The RMSD increased also with increasing hydration in methanol but was here less sensitive to hydration than in t-butanol. For both MTBE and hexane, the average RMSD seems to have a minimum at a hydration level between 150 and 200 and increases as the hydration level is further decreased. This is especially pronounced in hexane, where the RMSD becomes as high as 1.94 ± 0.03 Å at low hydration levels. Figures 7.9(a)–(e) are qualitatively similar to the RMSD plots for cutinase reported by Micaêlo and Soares (2007). In that study, the minimum RMSD was in hexane attained at a hydration level of 7.5% (w/w). The minima observed here for CALB in MTBE and hexane (Figures 7.9(d) and (e), respectively) correspond to a hydration level of 10–15%, which is similar to

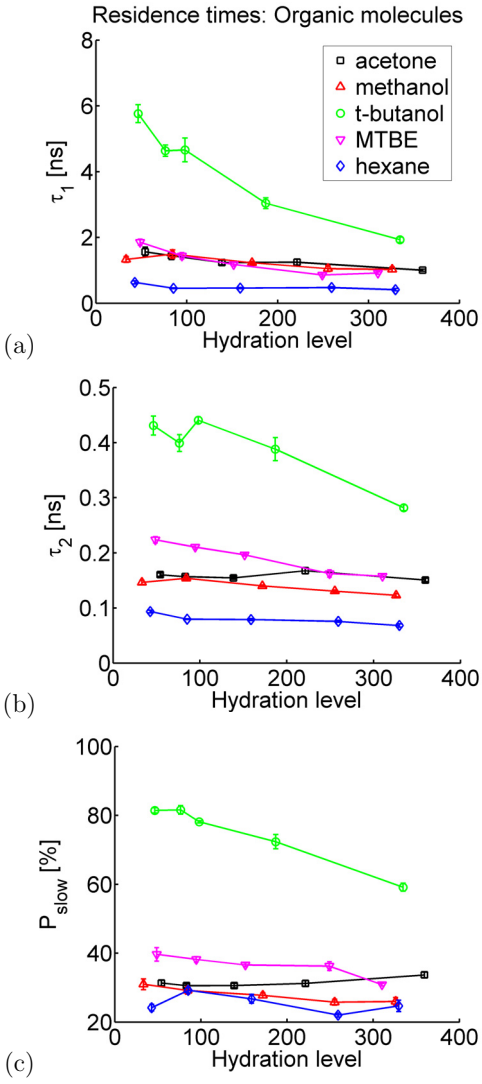


Figure 7.8: Residence times of (a) slowly and (b) rapidly exchanged organic solvent molecules vs. the hydration level determined in simulations with acetone (black squares + line), methanol (red triangles + line), t-butanol (green circles + line), MTBE (magenta triangles + line) or hexane (blue diamonds + line) as main solvent. In (c), the corresponding populations of slowly exchanged organic solvent molecules are shown. Standard errors estimates were based on three replica simulations which were started from different initial velocities.

the cutinase result.

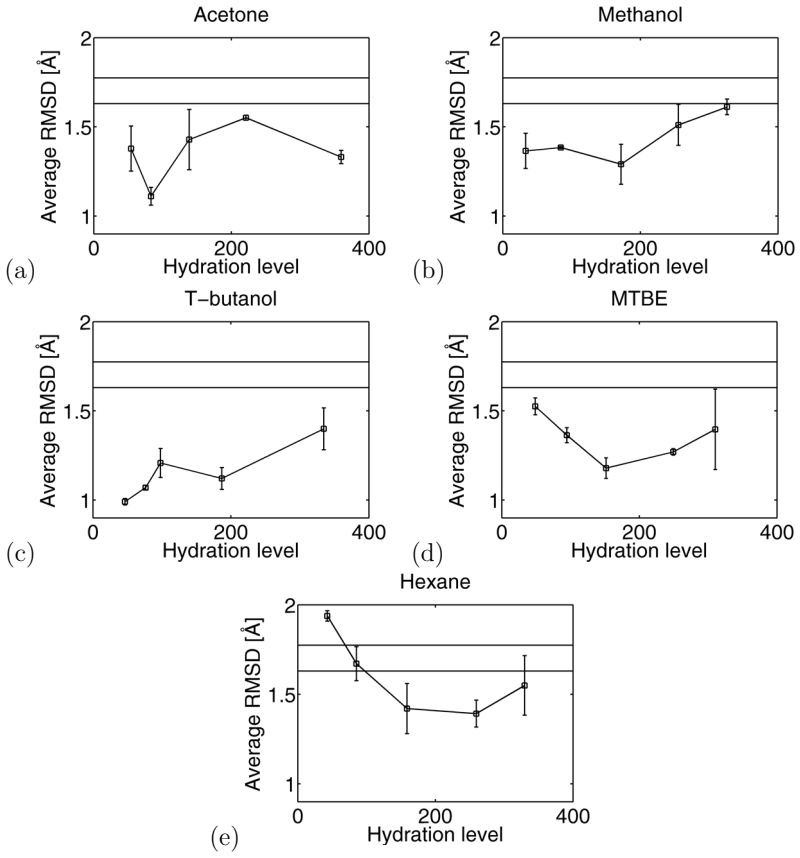


Figure 7.9: Average RMSD with respect to the crystal structure (1TCA) vs. hydration level for simulations carried out in (a) acetone, (b) methanol, (c) t-butanol, (d) MTBE and (e) hexane. Residues 1–20 and 308–317 were omitted from all RMSD evaluations. Standard error estimates were based on three replica simulations which were started from different initial velocities. Error bar limits for the average RMSD obtained from the pure water simulations is marked in each plot by horizontal lines.

The RMSD contribution from each residue was evaluated and for each simulation averaged over the final 10 ns. The results are not shown but briefly summarized below. For most individual residues and secondary structure elements, the RMSD seemed to be uncorrelated with solvent and hydration level. The trends in the total RMSD discussed above could therefore not be entirely attributed to the variation in RMSD of specific regions of CALB, but seemed to emerge when the entire protein was considered. The regions for which the local RMSD for some system significantly exceeded 2 Å were the loop L1 (residues 23–32), the helix α5 (residues 138–152), the loop L13 (residues 243–267) and the helix α10 (268–287). The high RMSD of L1 was caused by a slow fluctuating motion of the loop. The occurrence of this motion

seemed uncorrelated with solvent and hydration level. For the systems in which it occurred, the motion was typically only seen in one of the three replica simulations. For α5, the high RMSD was due to a partial or complete unfolding of the helix, as seen previously (see Chapter 3). This is discussed further in Section 7.3.2. The loop L13 showed an average RMSD greater than 2.5 Å in hexane and MTBE at low hydration levels (H43, H87, E50). As seen previously, the relatively high RMSD values of the loop were correlated with an elevated solvent accessible surface area (SASA) of the residue Tyr253 whose side chain was re-oriented and became solvent-exposed in those simulations. Simultaneously, the SASA of the negatively charged Asp252 decreased as the side chain was re-oriented towards the protein. The details of the SASA calculation are given in Section 7.3.3. The relatively high RMSD values observed for α10 in some simulations were caused by a partial un-winding of the helix. The unwinding was seen in simulations where methanol, hexane or pure water was the main solvent. The event did however not seem to be correlated with the hydration level. The unwinding occurred either in the part before the kink at Leu277 (residues 268–276) as seen previously (see Section 3.4.2), or in the part after the kink (residues 278–287) (the location of the kink is indicated in Figure 3.1). Simultaneous unwinding of both sides was not observed. In some of the hexane simulations, the helix was however seen to "straighten out", as Ala276 adopted Ramachandran angles in the helix region, and the kink at Leu277 seemingly disappeared.

7.3.2 Unfolding of α5

Similarly to the observations in Chapter 3, the helix α5 (residues 142–146) was seen to unfold to various extents in the different simulations. In three of the simulations carried out in t-butanol, T130(a), T130(c) and T210(a), the helical structure was maintained throughout the simulation as the backbone hydrogen bonds and the hydrogen bonds of Asp145/Ser150 and Asp145/Thr158 were intact. All remaining simulations essentially followed one of the five behaviors (A, B, C, D or E) described in Section 3.4.2 which was verified by visual inspection of the terminal frames of each simulation. The negatively charged Asp145 was either seen to hydrogen bond to Ser150 or Thr158 (situations D and C, respectively, see Figure 3.7(e) and (d)), to interact with Lys308 and Arg309 of the C terminal (situations A and B, see Figure 3.7(b) and (c)), to interact with Lys290 (not encountered in Chapter 3 but it resembles situation D) or to be directed into the solvent (situation E, see Figure 3.7(f)). In some cases, Asp145 got close enough to Ly290, Lys308 or Arg309 to form a salt bridge. As demonstrated previously (see Table 3.5), the cases A, B and C yielded the lowest values for the RMSD of residues 138–152 with respect to the crystal structure (1TCA). Case D yielded higher values while case E yielded the highest.

Table 7.2 summarizes which cases were observed in which solvents. It seems that the crystal structure conformation of α5 becomes less stable as solvent polarity or hydration level increases. Consider for instance case E, which is the case where α5 undergoes the largest structural changes (see Table 3.5). This occurred in all pure water simulations and half of the methanol simulations and was uncorrelated with the hydration level. It furthermore occurred in acetone, t-butanol and hexane, but only at hydration levels larger than 100.

Table 7.2: Summarizes which of the five different cases A, B, C, D and E (described in Section 3.4.2) were observed in the simulations of CALB. θ denotes here the hydration level.

Solvent	Cases observed
Acetone	B, C when $\theta < 100$
	E when $\theta > 100$
	D occurring at all θ
Methanol	D and E occurring at all θ
T-butanol	A, B, C occurring at all θ
	D, E when $\theta > 100$
	Helix intact in three simulations with $\theta < 100$
MTBE	Mainly A, B, C
Hexane	Mainly A, B, C
	E occurring in one simulation with $\theta = 330$
Water	Only E

For a quantitative analysis, the C_α atoms of residues 138–152 were for each frame aligned to the corresponding atoms in the crystal structure (1TCA), and the RMSD was evaluated. The obtained values were averaged over the last 10 ns of each simulation. The results, which are shown in Figures 7.10(a)–(e), support the above discussion. The RMSD of this region seems to increase with increasing solvent polarity. For acetone and t-butanol (Figures 7.10(a) and (c), respectively), it is also clear that the RMSD increases with increasing hydration.

7.3.3 Solvent-Accessible Surface Area

SASA calculations were carried out as described in Section 3.4.3, using VMD (Humphrey *et al.*, 1996). The obtained values were for each simulation averaged over the last 10 ns. The average total SASA of CALB is shown in Figures 7.11(a)–(e). For all solvents except methanol, the SASA clearly increased with increasing hydration. At a fixed hydration level, the values obtained in the different solvents were correlated with solvent polarity, with the smallest values in hexane and the largest in acetone. The SASA was generally lower in the organic solvents than in pure water. Water-like SASA values were nevertheless obtained in acetone at hydration levels larger than 150 and in methanol at all hydration levels. For all systems studied, the average total SASA was higher than that of the crystal structure (1TCA), except for hexane at low hydration (H43).

The SASA of hydrophobic and hydrophilic residues (as defined in Section 3.3) were as well evaluated for each system (data not shown). For methanol, neither the hydrophobic or hydrophilic SASA were correlated with hydration. For acetone, the hydrophilic SASA increased with increasing hydration while the hydrophobic SASA was unchanged. For t-butanol and MTBE, both the hydrophilic and hydrophobic SASA increased with increasing hydration at approximately the same rate. In hexane, a similar trend was observed although the hydrophilic SASA increased twice

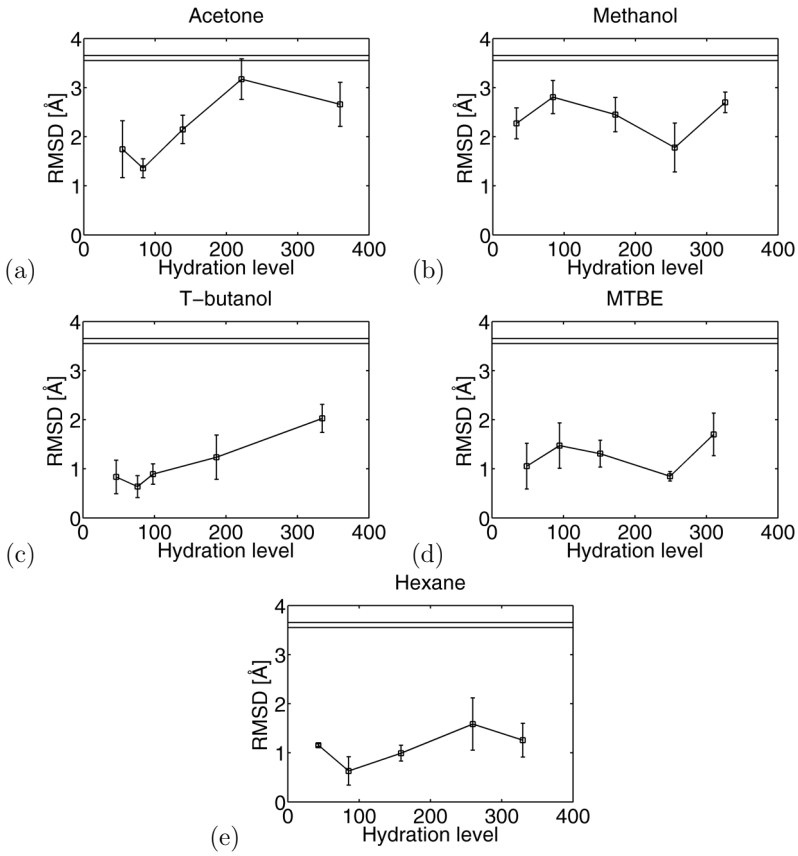


Figure 7.10: The average RMSD of C_α atoms of the $\alpha 5$ region (residues 138–152) obtained from simulations of CALB carried out in (a) acetone, (b) methanol, (c) t-butanol, (d) MTBE and (e) hexane and measured with the crystal structure (1TCA) as reference. Error bar limits for the average RMSD obtained from the pure water simulations are marked in each plot by horizontal solid lines.

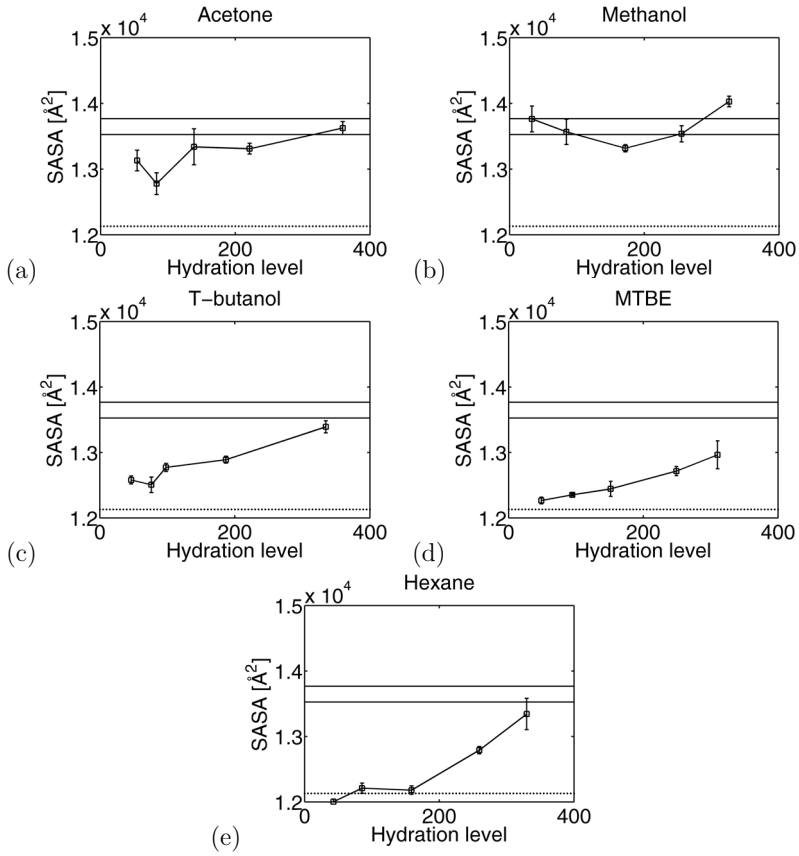


Figure 7.11: Average total SASA of CALB vs. the hydration level determined in simulations carried out in (a) acetone, (b) methanol, (c) t-butanol, (d) MTBE and (e) hexane. Standard error estimates were based on three replica simulations which were started from different initial velocities. Error bar limits for the average SASA obtained from the pure water simulations are marked in each plot by horizontal solid lines. SASA of crystal structure (1TCA) is marked with a horizontal dotted line.

as fast with increasing hydration as the hydrophobic. For all systems, the hydrophobic SASA was larger than in the crystal structure (1TCA). For H43, the hydrophilic SASA was 6500 ± 90 Å^2, which is significantly lower than the corresponding value for the crystal structure, 6980 Å^2 (Uppenberg *et al.*, 1994). In all other studied systems, the hydrophilic SASA was at least as large as the crystal structure value. This demonstrates that particularly in hexane, there was a tendency for the hydrophilic surface area to be reduced as the hydration level decreased. This accounted for the low total SASA seen in hexane (Figure 7.11(e)) and possibly also for the increase in RMSD measured from the crystal structure (1TCA) discussed in Section 7.3.1 (see also Figure 7.9(e)).

For each system, the average exposed surface area fraction was evaluated for each protein residue. This quantity is defined as the ratio between the SASA for the residue as measured in the current protein conformation and the SASA of the residue evaluated as if the rest of the protein was transparent. The results are not shown here but were employed in the calculation of water and organic solvent residence times described in Section 7.2.4.

7.4 Flexibility

The flexibility of CALB was assessed by calculation of B-factors, which were defined in Section 3.5. For each simulation, B-factors for each C_α atom were evaluated based on the last 10 ns of simulation.

In order to identify regions of high flexibility and to qualitatively assess how the solvent impacts these regions, the B-factors of the C_α atoms were for each organic solvent averaged over the six individual simulations corresponding to the two lowest hydration levels. For water, the B-factors were averaged over the five replica simulations W(a)–(e). The results are shown in Figures 7.12(a)–(b). The flexibility was in some cases high in the regions consisting of residues 23–32 (L1), 138–152 (α5 and adjacent loop section), 184–207 (L11), 243–267 (L13) and 268–287 (α10). This is consistent with the results of Section 3.5.

In order to characterize the overall flexibility of CALB, average B-factors β_{av} were calculated for each simulation by averaging over all C_α atoms excluding the N- and C-termini (residues 1–20 and 308–317). The results are shown in Figures 7.13(a)–(e). Due to the spread of the obtained values, any correlation between $\ln \beta_{\text{av}}$ and the hydration level could not be observed by mere visual inspection. The approach was therefore taken to fit the model expression

$$\ln \beta_{\text{av}} = p_0 + p_1 \theta \tag{7.6}$$

to the values of $\ln \beta_{\text{av}}$ obtained in each organic solvent. θ denotes here the hydration level and p_0 and p_1 are adjustable parameters. $\ln \beta_{\text{av}}$ was used rather than β_{av} since this yielded residuals more closely following a Gaussian distribution leading to a more reliable statistical analysis. The regressed models are shown in Figures 7.13(a)–(f), and the values of the parameters p_0 and p_1 are listed in Table 3.7 along with estimates of $\ln \beta_{\text{av}}$ at $\theta = 200$ and 400. Two-sided 95% confidence intervals were evaluated under the assumption that the residuals were statistically independent and followed a Gaussian distribution. The parameter p_1 which gives the dependence

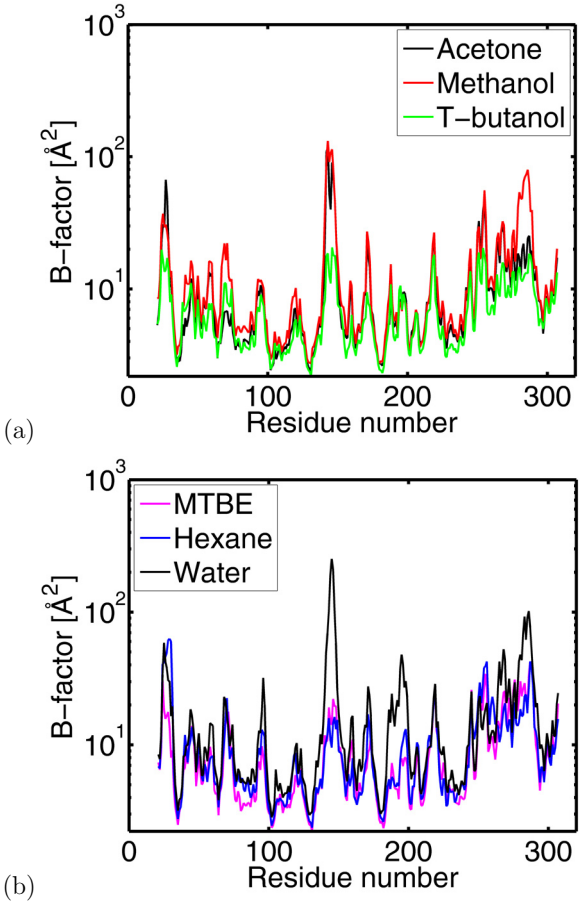


Figure 7.12: B-factors for C_α atoms obtained from simulations carried out in acetone (black), methanol (red) and t-butanol (green) are shown in (a). In (b), B-factors obtained from MTBE (magenta), hexane (blue) and pure water (black) are shown. For each organic solvent, values shown are averages over the six individual simulations corresponding to the two lowest hydration levels. For pure water, values shown are averages over the five simulations W(a)–(e). Values for N- and C-termini (residues 1–20 and 308–317) are omitted.

of $\ln \beta_{av}$ on the hydration level was significantly different from zero for all solvents except methanol. The obtained values for all solvents indicated that $\ln \beta_{av}$ increased with increasing hydration. The parameter p_0 gives $\ln \beta_{av}$ in the current solvent at a hydration level of zero, assuming that Equation (7.6) is valid in this limit. This parameter, along with the estimates at $\theta = 200$ and 400 can therefore be taken as an approximate measures of the extent to which organic solvent promotes flexibility. The values of p_0 and $\ln \beta_{av}$ at $\theta = 200$ ranked the solvents in the order of increasing polarity, with the striking exception that the lowest value was obtained in t-butanol.

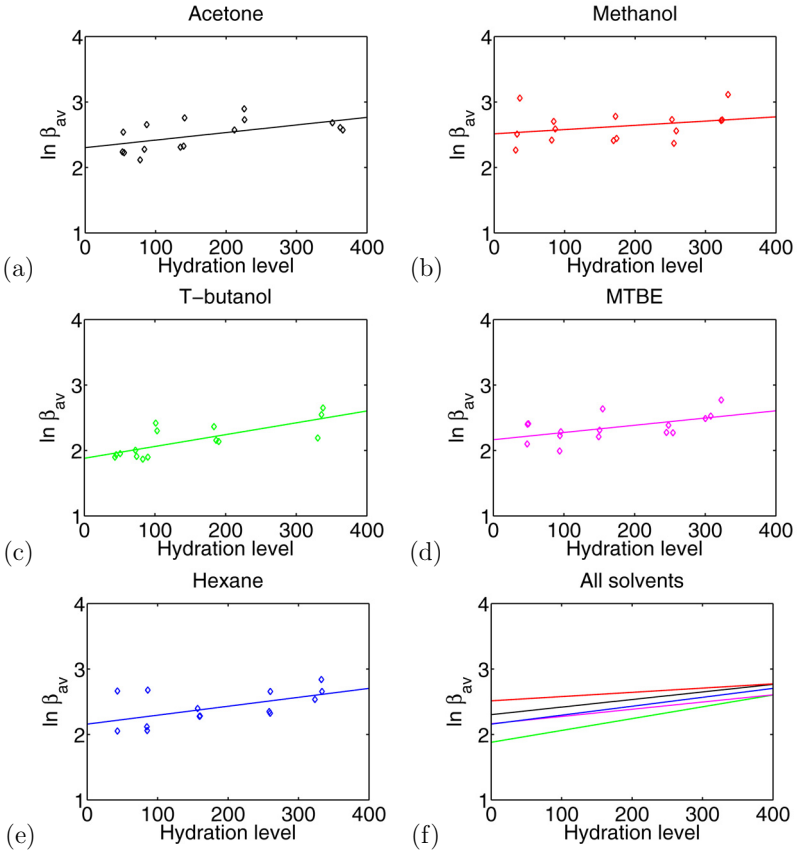


Figure 7.13: Values of $\ln \beta_{av}$ obtained from simulations carried out in (a) acetone, (b) methanol, (c) t-butanol, (d) MTBE and (e) hexane vs. hydration level. For each solvent, Equation (7.6) has been fitted to the data and the results are shown as lines. In (f), the models corresponding to the different solvents are compared.

The same procedure was employed to investigate correlations between the B-factors of the high flexibility regions, as given above, and the hydration level. For the L11 region, the average B-factor increased with increasing hydration for all five organic solvents. For the other regions, the B-factors were too dispersed and significant correlations could not be established. These results are not shown.

Table 7.3: The parameters p_0 and p_1 obtained from fitting Equation (7.6) to values of $\ln \beta_{\mathrm{av}}$ obtained in the different solvents. Also given are estimates of $\ln \beta_{\mathrm{av}}$ at $\theta = 200$ and 400, where θ denotes the hydration level, based on the regressed models. The value given for pure water is the average of values of $\ln \beta_{\mathrm{av}}$ obtained from the simulations W(a)–(e). The error intervals are 95% two-sided confidence intervals for the parameters p_0 and p_1 derived from the linear regression.

			$\ln \beta_{\mathrm{av}}$	
Solvent	$p_1 \times 10^3$	$\theta = 0$ (p_0)	$\theta = 200$	$\theta = 400$
Acetone	1.2 ± 1.0	2.30 ± 0.21	2.53 ± 0.12	2.76 ± 0.26
Methanol	0.6 ± 1.2	2.51 ± 0.26	2.64 ± 0.14	2.77 ± 0.31
T-butanol	1.8 ± 0.9	1.88 ± 0.17	2.24 ± 0.11	2.60 ± 0.25
MTBE	1.1 ± 1.0	2.16 ± 0.20	2.38 ± 0.10	2.61 ± 0.25
Hexane	1.4 ± 1.2	2.16 ± 0.24	2.43 ± 0.13	2.70 ± 0.29
Water	-	2.82 ± 0.33	-	

In order to investigate how B-factors in these regions depended on the organic solvent, the average B-factors of the regions were for each solvent further averaged over the six simulations corresponding to the two lowest hydration levels (which for all solvents were less than 100). The results, which are given in Table 7.4, demonstrate that the organic solvent had a significant impact on the flexibility of the selected regions. The trends were however different for the different regions. α5 became more flexible as the polarity of the solvent increased with the exception that very low flexibility was obtained in t-butanol. The average B-factor of L11 was low in all five organic solvents, but significantly higher in pure water. For L13, the flexibility was lower in water than in any of the organic solvents, again with the exception of t-butanol. For α10, the flexibility was highest in water and methanol, slightly lower in hexane and MTBE and as lowest in acetone and t-butanol. For acetone, hexane and pure water, these trends were fully consistent with the results reported in Chapter 3.

These results indicate that the flexibility of CALB in organic media generally increases with increasing hydration. This is somewhat different from previous simulations of CALB in gaseous water/argon mixtures, in which no significant increase in flexibility was detected (Branco *et al.*, 2009). The results confirm the previous finding that the flexibility depends on the bulk organic solvent (Trodler and Pleiss, 2008). The role of the solvent for modulating flexibility seems however to be more complex than that of merely being a medium with a certain ability to dissolve the water layer around the protein. It is here demonstrated that the overall flexibility of CALB is different in different solvents, even if compared at similar hydration levels, which corresponds to approximately similar water activities, as shown in Section 7.2.2. Important for the flexibility is possibly that the organic solvent modulates the dynamical behavior of the water molecules bound to the protein. The rapidly exchanged water molecules probably promote protein flexibility while the slowly exchanged water molecules counteract it, as argued by (Trodler and Pleiss, 2008). According to Section 7.2.4, the fraction of slowly exchanged water molecules is larger

Table 7.4: B-factors averaged over selected regions of CALB. Values listed for each solvent are averages of the six individual simulations corresponding to the two lowest hydration levels. For water, listed values are averages over the five simulations W(a)–(e). Error intervals are regular standard error estimates.

	Average B-factor in region [Å^2]			
Solvent	α5 (138–152)	L11 (184-207)	L13 (250–256)	α10 (268–287)
Acetone	42 ± 18	6.1 ± 0.2	31 ± 2	18 ± 2
Methanol	53 ± 20	6.6 ± 0.3	31 ± 3	39 ± 10
T-butanol	12 ± 1	5.8 ± 0.3	16 ± 3	12 ± 2
MTBE	18 ± 4	5.6 ± 0.3	40 ± 12	22 ± 3
Hexane	11 ± 1	6.9 ± 0.6	31 ± 17	19 ± 3
Water	76 ± 21	18 ± 1	18 ± 1	46 ± 17

in hexane, MTBE and t-butanol than in acetone and methanol. The residence times of slow water molecules are furthermore higher in the three former solvents. This could explain the lower flexibility observed in the three former solvents.

To explain that the lowest flexibility of CALB was encountered in t-butanol, it seems however necessary to consider direct interactions between the protein and organic solvent molecules. Solvents that have a polar group like acetone, methanol and t-butanol might bind to the protein surface mimicking hydration water molecules. It was however shown in Section 7.2.4 via analysis of residence times that t-butanol molecules bind much tighter to the protein surface than acetone and methanol molecules do. Thus, a t-butanol molecule might to a much larger extent play the role as a slowly exchanged water molecule and thus counteract the flexibility of CALB. MTBE and hexane would not restrict the flexibility as much as t-butanol as these molecules do not bind the protein as tight, due to the absence of polar groups, and consequently have shorter residence times than t-butanol molecules.

7.5 Summary

CALB has in this chapter been studied in pure water and the organic solvents acetone, methanol, t-butanol, MTBE and hexane at several hydration levels, by MD simulation. For the simulations carried out in the three polar organic solvents, the bulk water activity was evaluated. At similar activities, the number of water molecules in the first solvation shell, i.e. the hydration level, was approximately the same in the three solvents. Analysis of the water clusters on the surface of CALB showed that at low a_w, water molecules bind to the surface individually or in small clusters. At $a_w \approx 0.4$–0.5, the water molecules start binding to already present water clusters, which start to percolate, consistent with a previous result of Branco *et al.* (2009). Calculation of residence times of water molecules at individual protein residues revealed that hydration water molecules became more volatile with increasing hydration level and with increasing solvent polarity, which is consistent with previous studies (Micaêlo and Soares, 2007).

The key conclusion of this chapter is nevertheless that structural and dynamical properties of CALB such as RMSD, SASA and B-factors are sensitive to the hydration level. Increasing the hydration level did however have different impacts in the different solvents. The average total RMSD measured from the crystal structure had for instance a U-shaped dependence on the hydration level in MTBE and hexane, but increased monotonically with increasing hydration in t-butanol. The overall flexibility of CALB was found to increase with increasing hydration level in all solvents except for methanol. At similar hydration levels, the flexibility increased with increasing solvent polarity, with the exception that the lowest flexibility was observed in t-butanol. A possible explanation for these trends in terms of the residence times of water and organic solvent molecules was given.

8

Conclusion

8.1 Summary

Conventional molecular design strategies are limited for the task of selecting solvent for biocatalytic processes to be carried out in non-aqueous media. This is due to that molecular interactions between the enzyme and its solvent environment are complex and are not yet fully understood. Within this work, an increased understanding of enzymes in non-aqueous media was sought by means of molecular dynamics (MD) simulations. The investigations focused on the enzyme *Candida antarctica* lipase B (CALB) in mixtures of organic solvent and water. Special attention was given to how the water was distributed in the system, and how structure and dynamics of CALB were dependent on the number of water molecules present. In particular, an approach was developed to calculate the thermodynamic water activity of the medium, and it was applied to investigate the significance of water activity for properties of CALB.

The work comprised two simulations studies of CALB. The first study (Chapter 3) considered CALB at 25 °C in the solvents acetone, hexane and pure water and comprised in total 9 simulated systems. For the two organic solvents, several systems were simulated including different numbers of water molecules. The hydration level of CALB was measured as the number of water molecules located in the first solvation shell of the enzyme. Both the solvent and the hydration level were shown to have a measurable effect on the structure and dynamics of CALB. The solvent-accessible surface area, for instance, increased with increasing hydration, and as well with solvent polarity. The flexibility, which was measured by the B-factors, also increased with increasing hydration. Certain regions of CALB were identified, for which the structure and flexibility were especially sensitive to solvent and hydration level. In particular, the simulations confirmed the previous observation that an α-helix located on the rim of the active site pocket unfolds, and might function similarly to a lid (Skjøt *et al.*, 2009). It was furthermore shown here that the structural changes of this helix depend on the solvent and the hydration level, as the smallest changes measured from the crystal structure were seen in dry acetone or hexane, while the largest changes were seen in pure water.

The second study (Chapter 7) considered CALB at 50 °C in the organic solvents acetone, methanol, t-butanol, methyl t-butyl ether (MTBE) and hexane, which are commonly used in non-aqueous biocatalysis and span a broad range of solvent polarities. The enzyme was also studied in pure water and the study comprised 26 simulated systems in total. For each organic solvent, five different hydration levels were considered. For acetone, methanol and t-butanol, the thermodynamic water activity of the systems was determined. Adsorption isotherms were determined con-

sidering the water molecules in the first solvation shell as adsorbed, and were found to be nearly independent of the solvent. This confirmed the common assumption that similar hydration levels are obtained at similar water activities (Halling, 1994). The structure and flexibility of CALB were significantly dependent on the hydration level in all of the organic solvents except for methanol. A consequence of this is that the hydration level needs to be carefully considered if one is interested in comparing the effects of different organic solvents. The solvent was also found to impact the structure and flexibility of CALB. Comparisons made at similar hydration levels showed that the flexibility correlated with solvent polarity, with the exception that lowest flexibility was observed in t-butanol. Dry t-butanol was also found to be the medium best stabilizing the "lid-candidate" helix.

A significant part of the work was the development of the method for determining the water activity (Chapter 4). In the approach taken and here termed *a posteriori* analysis, the fraction of water molecules in the bulk medium, i.e. far from the protein surface, is determined in the protein simulation. The activity corresponding to this fraction is obtained from an excess Gibbs energy model for the corresponding water/organic solvent mixture. In order to obtain these excess Gibbs energy models, a previously developed methodology based on fluctuation solution theory (FST) was employed (Christensen *et al.*, 2007a). In this approach, simulations are carried out of the binary mixture at several compositions. At each composition, the pair radial distribution functions (RDFs) are spatially integrated, and the obtained integrals, i.e. the total correlation function integrals (TCFIs), are used to calculate activity coefficient derivatives for the mixture. The modified Margules model for the excess Gibbs energy is then fitted to these derivatives.

Since the TCFIs rarely converge within the range over which the RDFs are sampled, assumptions need to be made of the long-range behavior of the RDFs, in order to obtain accurate results. Previously suggested approaches were found to be inadequate for the present mixtures, since water was one of the components. Therefore, attempts were made to develop a robust and theoretically motivated method for predicting the long-range behavior of RDFs. A method used by Verlet (1968) to extend the RDF of the pure Lennard-Jones fluid was applied and extended for application to molecular fluid mixtures. In this method, the direct correlation functions (DCFs) of the mixture are assumed to follow a certain approximate expression at long distances. By numerical solution of the Ornstein-Zernike (OZ) equation, which relates the DCFs and the RDFs, the corresponding long-range behavior of the RDFs is obtained, and the integral can be extended until convergence. In this work, an approximation for the DCF at long distances was derived for molecular fluids with interactions described by the CHARMM force field, for rigid molecules of zero ionic strength (Chapter 5).

An extensive set of simulations was carried out in order to validate the extended Verlet method (Chapter 6). The studies progressed by considering pure atomic fluids, binary atomic mixtures, pure molecular fluids and finally binary molecular mixtures. The TCFIs from the extended RDFs were validated by comparing the results for thermodynamic derivatives, such as isothermal compressibilities, partial molecular volumes and activity coefficient derivatives with properties obtained from alternative computational routes or from equations of state fitted to previous simulations. The extended Verlet method was found to yield properties of good accuracy,

except in the case of systems near the critical point or mixtures at compositions where one component is dilute. The method was furthermore shown to yield results of at least as good accuracy as two previously proposed methods for obtaining TCFIs from molecular simulations, namely the methods of Weerasinghe and Smith (2003) and Hess and van der Vegt (2009).

8.2 Contribution of this PhD Thesis

The most significant achievements of this work are summarized below.

- The structure and flexibility of CALB in several organic solvents have been shown to depend on the hydration level by MD simulations. This implies that simulation studies investigating the effects of different organic solvents need to consider the enzyme hydration carefully, in order to reach correct conclusions.
- A methodology has been developed for determining the water activity in simulation of enzymes in water-miscible organic solvents. This allows for comparisons of enzyme properties in different organic solvents to be made at similar water activities.
- The water adsorption isotherms of CALB were shown to be nearly identical in acetone, methanol and t-butanol. This seems to be the first simulation study that explicitly gives support to the common assumption that similar hydration level is obtained at similar water activity in different organic solvents (Halling, 1994).
- The method of Verlet (1968) for extending RDFs obtained from simulations have been extended to the case of molecular fluid mixtures, and it has been shown to yield accurate TCFIs for a broad range of systems at different state conditions and mixture compositions, and for molecules as anisotropic as hexane. The main advantage over previously proposed methods for calculating TCFIs is that by using the Verlet method, accurate results can be obtained by simulations of smaller systems. This might allow for accurate prediction of thermodynamic properties of mixtures with relatively small computational efforts.

8.3 Future Work

Several significant tasks were not accomplished within this work and are left to future investigations. Important considerations are summarized below.

- An important extension would be to develop a method for determining the water activity in a protein simulation in which the organic solvent is immiscible with water. According to Chapter 7, the bulk fraction of water molecules in MTBE and hexane can very well be determined by simulations comprising 20 ns. The corresponding activity coefficients can however not be obtained by the FST methodology. Possible alternative approaches could be based on Gibbs-ensemble Monte Carlo, or free energy perturbation simulations. In this work,

the hydration level was found to be a nearly universal function of the water activity in water-miscible solvents. Investigating whether this result extends to water-immiscible solvents would be of significance.

- One of the bottlenecks in MD studies of enzymes in organic media is that the availability of accurate force field parameters is limited for common industrial solvents. It was for instance shown in Chapter 6 that with the current CHARMM parameters for acetone, the thermodynamics of water/acetone mixtures is in poor agreement with experimental findings. Establishing principles and optimization methods for systematic parameterization of new solvent molecules would increase the predictive power of MD simulations for non-aqueous biocatalysis.

- The purpose of simulation studies such as this is ultimately to investigate if molecular interactions can explain solvent effects on enzyme properties such as activity, specificity and stability. Correlating simulation results with experimental data would be facilitated if new, more systematic experiments are performed. Such experiments should measure mentioned properties in a range of solvents, at controlled water activities and for different substrates and reactions. The results would guide planning and analysis of MD simulations which would explore if experimental trends can be attributed to molecular phenomena.

- The extended Verlet method is based on an approximate OZ equation (Equation (5.8)) which neglects the coupling between the isotropic and anisotropic correlation functions. Although this has not been shown to limit the performance of the method, it would be worthwhile to test the significance of the assumption more explicitly. This can be done as outlined in Section 5.2, by representing the correlation functions by truncated spherical harmonics expansions (Gray and Gubbins, 1984).

A

The CHARMM Force Field and Parameters

The CHARMM force field employs the following expression for the potential energy (MacKerell Jr. *et al.*, 1998)

$$\begin{aligned} U_{\text{total}} = & \\ & \sum_{\text{bond}} K_b \left(b - b_0\right)^2 + \sum_{\text{angle}} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{UB}} K_{\text{UB}} \left(S - S_0\right)^2 + \\ & \sum_{\text{dihedral}} K_\chi \left(1 + \cos\left(n\chi - \delta\right)\right) + \sum_{\text{improper}} K_\phi \left(\phi - \phi_0\right)^2 + \\ & \sum_{\text{nonbond}} \left(\epsilon_{ij} \left[\left(\frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\text{min},ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon_1 r_{ij}} \right) \end{aligned} \tag{A.1}$$

where K_b, K_θ, K_{UB}, K_χ and K_ϕ denote respectively bond, angle, Urey-Bradley, dihedral angle and improper dihedral angle force constants. The constants b, θ, S, χ and ϕ denote respectively bond length, bond angle, Urey-Bradley 1,3-distance, dihedral angle and improper torsion angle. The subscript zero denotes the corresponding equilibrium values for the individual terms. The constants n and δ denote respectively the symmetry number and phase shift for the dihedral angle energy term. For the non-bonded energy term, r_{ij} denotes the distance between atoms i and j. The parameters ϵ_{ij} and $R_{\text{min},ij}$ denote respectively the Lennard-Jones potential well depth and equilibrium distance. For unlike atom types, these parameters are calculated with the Lorentz-Berthelot combining rules

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \tag{A.2}$$

$$R_{\text{min},ij} = \frac{R_{\text{min},i} + R_{\text{min},j}}{2} \tag{A.3}$$

The parameters q_i and ϵ_1 denote respectively the partial charge of atom i and the effective dielectric constant, which is set to unity throughout this work.

Figures A.1(a)–(e) show the molecular topologies of the organic molecules studied in this work, indicating CHARMM atom types for each atom. Tables A.1–A.5 respectively list partial charges, Lennard-Jones, bond, angle and dihedral parameters used in this work for the organic solvent molecules. Impropers were not used for any of these molecules.

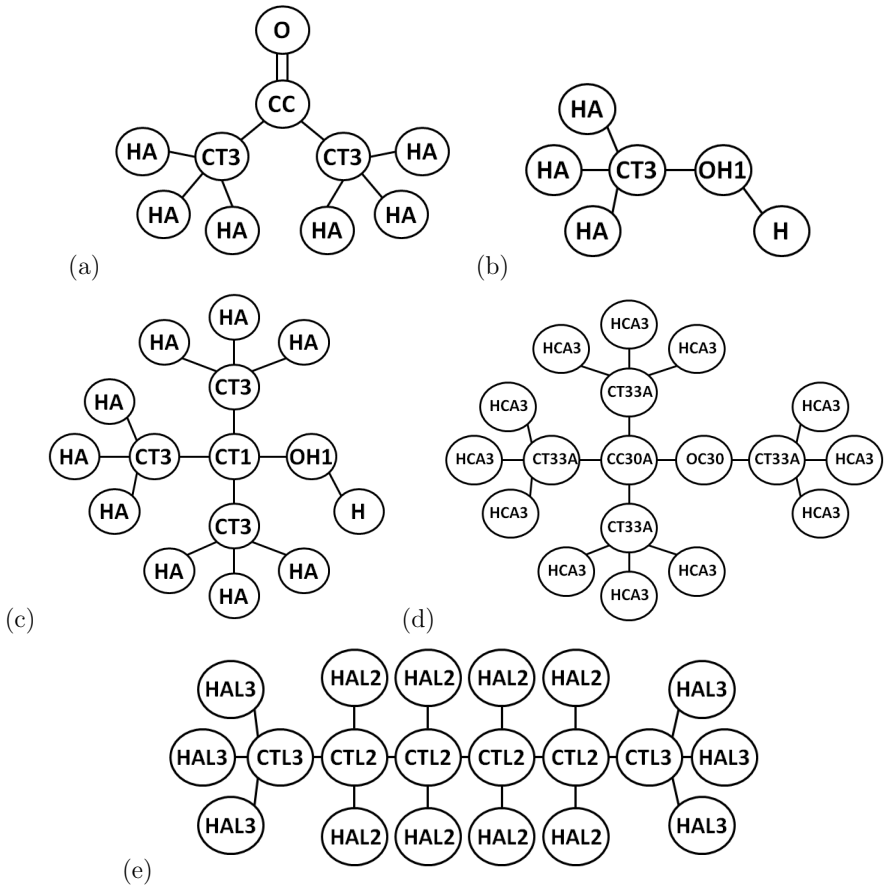


Figure A.1: Topology diagrams for the all-atom representation of (a) acetone, (b) methanol, (c) t-butanol, (d) MTBE and (e) hexane, indicating the CHARMM atom types.

Table A.1: Atom types and partial charges taken from the CHARMM27 (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) and CHARMM35 (ethers) (Vorobyov *et al.*, 2007) force fields, and CHARMM parameters used for acetone by Martin and Biddy (2005). *Missing parameter. Reported values were taken from parameters for similar atom types.

Atom	Description	q
Acetone		
CT3	methyl carbon	-0.27
CC	ketone carbon	0.55
O	ketone oxygen	-0.55
HA	methyl hydrogen	0.09
Methanol		
CT3	methyl carbon	-0.04
OH1	alcohol oxygen	-0.66
HA	methyl hydrogen	0.09
H	alcohol hydrogen	0.43
T-butanol		
CT3	methyl carbon	-0.27
CT1*	tertiary carbon	0.23
OH1	alcohol oxygen	-0.66
HA	methyl hydrogen	0.09
H	alcohol hydrogen	0.43
MTBE		
CC33A	methyl carbon	-0.27
CC33A	ether-bonded methyl carbon	-0.10
CC30A*	ether-bonded tertiary carbon	0.17
OC30A	ether oxygen	-0.34
HCA3	methyl hydrogen	0.09
Hexane		
CTL3	methyl carbon	-0.27
CTL2	methylene carbon	-0.18
HAL3	methyl hydrogen	0.09
HAL2	methylene hydrogen	0.09

Table A.2: Lennard-Jones parameters taken from the CHARMM27 (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) and CHARMM35 (ethers) (Vorobyov *et al.*, 2007) force fields, and CHARMM parameters used for acetone by Martin and Biddy (2005).

Atom type	ϵ [kcal/mole]	$R_{\min}/2$ [Å]	ϵ^{1-4} [kcal/mole]	$R_{\min}^{1-4}/2$ [Å]
CT3	-0.0800	2.0600	-0.01	1.9
CT1	-0.0200	2.2750	-0.01	1.9
CC	-0.0700	2.0000	-	-
CTL3	-0.0780	2.0400	-0.01	1.9
CTL2	-0.0560	2.0100	-0.01	1.9
CC33A	-0.0780	2.0400	-0.01	1.9
CC30A	-0.0320	2.0000	-0.01	1.9
OH1	-0.1521	1.7700	-	-
O	-0.1200	1.7000	-0.12	1.4
HA	-0.0220	1.3200	-	-
H	-0.0460	0.2245	-	-
HAL3	-0.0240	1.3400	-	-
HAL2	-0.0280	1.3400	-	-
HCA3	-0.0240	1.3400	-	-
OC30A	-0.1000	1.6500	-	-

Table A.3: Bond parameters taken from the CHARMM27 (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) and CHARMM35 (ethers) (Vorobyov *et al.*, 2007) force fields, and CHARMM parameters used for acetone by Martin and Biddy (2005). *Missing parameter. Reported values were taken from parameters for similar atom types.

Bond	K_b [kcal/mole]	b_0 [Å]
CT3–CT1	222.5	1.538
CT3–CC	200.0	1.522
CT3–OH1	428.0	1.420
CT3–HA	322.0	1.111
CT1–OH1	428.0	1.420
CC–O	650.0	1.230
OH1–H	545.0	0.960
CC33A–CC30A	222.5	1.538
CC33A–OC30A	360.0	1.415
CC33A–HCA3	322.0	1.111
CC30A–OC30A*	360.0	1.415
CTL3–CTL2	222.5	1.528
CTL3–HAL3	322.0	1.111
CTL2–CTL2	222.5	1.530
CTL2–HAL2	309.0	1.111

Table A.4: Angle parameters taken from the CHARMM27 (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) and CHARMM35 (ethers) (Vorobyov *et al.*, 2007) force fields, and CHARMM parameters used for acetone by Martin and Biddy (2005). *Missing parameter. Reported values were taken from parameters for similar atom types.

Angle	K_θ [kcal/mole]	θ_0	K_{UB} [kcal/mole]	S_0 [Å]
CT3–CT1–OH1	75.70	110.1	-	-
CT3–CT1–CT3	53.35	114.0	8.00	2.561
CT1–OH1–H	57.50	106.0	-	-
CT1–CT3–HA	33.43	110.1	22.53	2.179
HA–CT3–HA	35.50	108.4	5.40	1.802
CT3–OH–H	57.50	106.0	-	-
OH1–CT3–HA	45.90	108.9	-	-
CT3–CC–CT3	50.00	116.5	50.00	2.450
CT3–CC–O	15.00	121.0	50.00	2.440
CC–CT3–HA	33.00	109.5	30.00	2.163
CC33A–CC30A–CC33A	53.30	114.0	8.00	2.561
CC33A–CC30A–OC30A*	45.00	111.5	-	-
CC30A–CC33A–HCA3	33.43	110.1	22.53	2.179
CC30A–OC30A–CC33A*	95.00	109.7	-	-
OC30A–CC33A–HCA3	60.00	109.5	-	-
HCA3–CC33A–HCA3	35.50	108.4	5.40	1.802
CTL3–CTL2–CTL2	58.00	115.0	8.00	2.561
CTL3–CTL2–HAL2	34.60	110.1	22.53	2.179
CTL2–CTL3–HAL3	34.60	110.1	22.53	2.179
CTL2–CTL2–CTL2	58.35	113.6	11.16	2.561
CTL2–CTL2–HAL2	26.50	110.1	22.53	2.179
HAL3–CTL3–HAL3	35.50	108.4	5.40	1.802
HAL2–CTL2–HAL2	35.50	109.0	5.40	1.802

Table A.5: Dihedral angle parameters taken from the CHARMM27 (MacKerell Jr. *et al.*, 1998; MacKerell Jr. *et al.*, 2004) and CHARMM35 (ethers) (Vorobyov *et al.*, 2007) force fields, and CHARMM parameters used for acetone by Martin and Biddy (2005). *Missing parameter. Reported values were taken from parameters for similar atom types.

Dihedral	K_χ [kcal/mole]	n	δ
CT3–CT1–OH1–H	1.33	1	0
CT3–CT1–OH1–H	0.18	2	0
CT3–CT1–OH1–H	0.32	3	0
CT3–CT1–CT3–HA	0.20	3	0
OH1–CT1–CT3–HA	0.20	3	0
CT3–CC–CT3–HA	0.05	6	180
O–CC–CT3–HA	0.05	6	180
HA–CT3–OH1–H	0.14	3	0
CC33A–CC30A–CC33A–HCA3*	0.160	3	0
CC33A–CC30A–OC30A–CC33A*	0.400	1	0
CC33A–CC30A–OC30A–CC33A*	0.490	3	0
CC30A–OC30A–CC33A–HCA3*	0.284	3	0
OC30A–CC30A–CC33A–HCA3*	0.160	3	0
CTL3–CTL2–CTL2–CTL2	0.10	2	180
CTL3–CTL2–CTL2–CTL2	0.19	3	0
CTL3–CTL2–CTL2–CTL2	0.15	4	0
CTL3–CTL2–CTL2–CTL2	0.10	6	180
CTL3–CTL2–CTL2–HAL2	0.19	3	0
CTL2–CTL2–CTL3–HAL3	0.16	3	0
CTL2–CTL2–CTL2–CTL2	0.10	2	180
CTL2–CTL2–CTL2–CTL2	0.19	3	0
CTL2–CTL2–CTL2–CTL2	0.15	4	0
CTL2–CTL2–CTL2–CTL2	0.10	6	180
CTL2–CTL2–CTL2–HAL2	0.19	3	0
HAL3–CTL3–CTL2–HAL2	0.16	3	0
HAL2–CTL2–CTL2–HAL2	0.19	3	0

B

RMSD Plots for CALB Study I

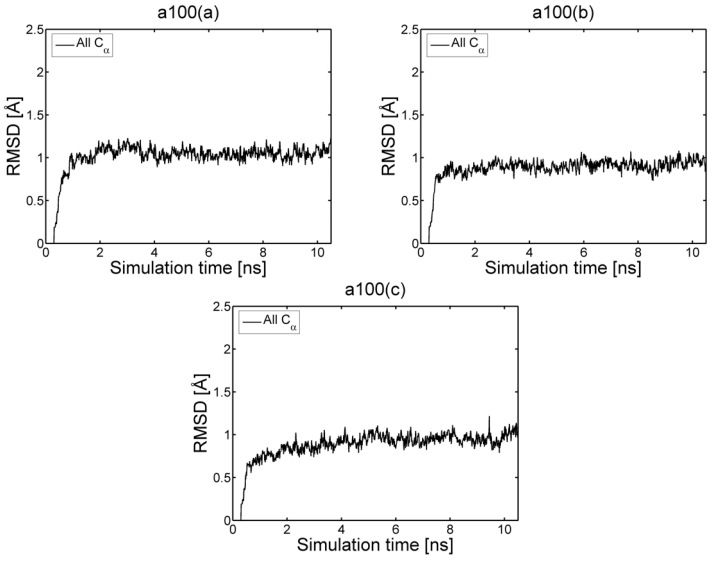


Figure B.1: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 3.2. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 1–10 (R1), 23–32 (R2), 138–152 (R3), 190–202 (R4), 243–292 (R5) and 308–317 (R6).

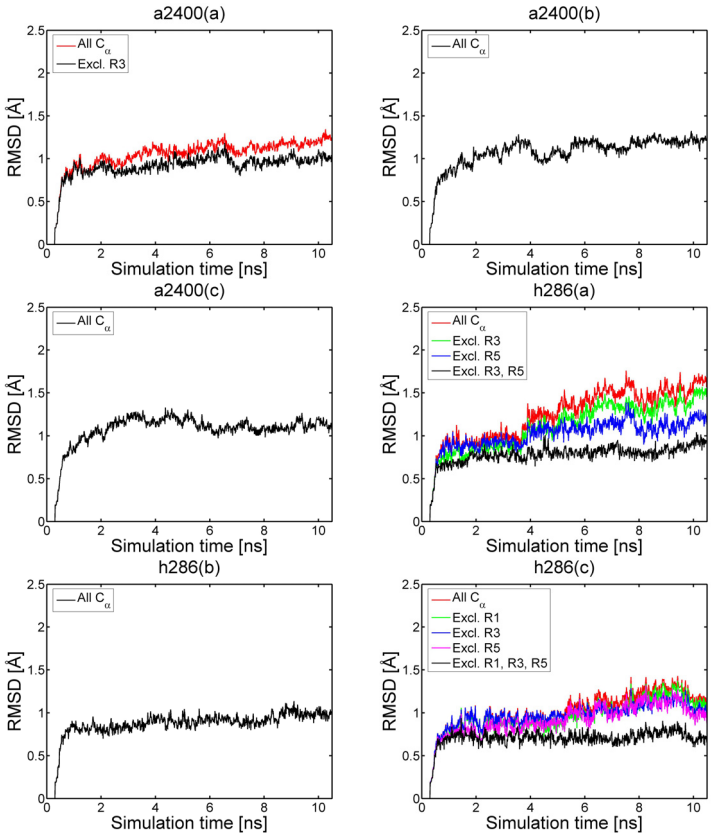


Figure B.2: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 3.2. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 1–10 (R1), 23–32 (R2), 138–152 (R3), 190–202 (R4), 243–292 (R5) and 308–317 (R6).

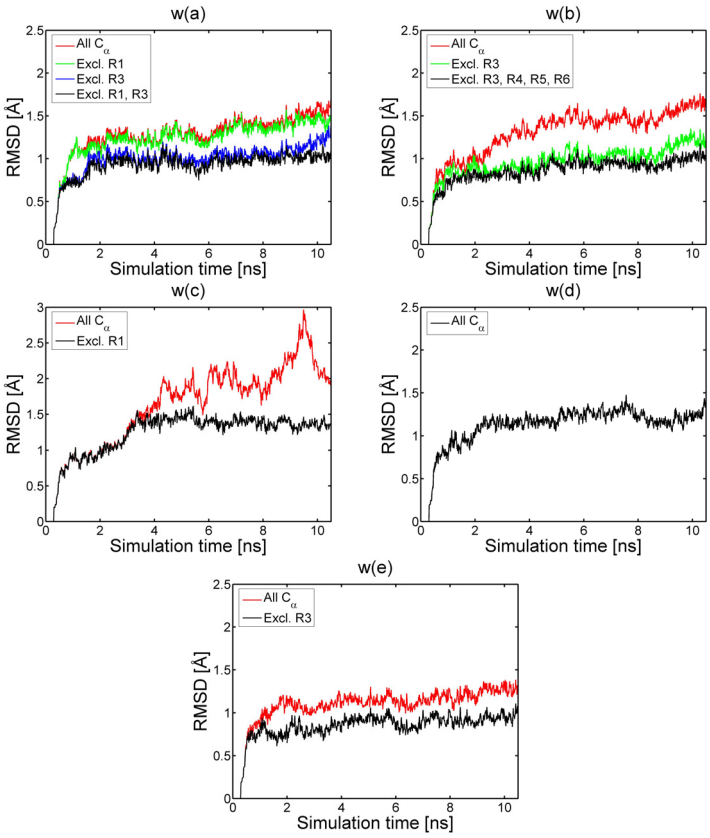


Figure B.3: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 3.2. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 1–10 (R1), 23–32 (R2), 138–152 (R3), 190–202 (R4), 243–292 (R5) and 308–317 (R6). Observe that the y-scale for w(c) is different.

C

Jacobians for Solution of the Verlet Method Equations

In this appendix, the notation follows that of Section 5.3.2, with exception for that the superscript "(t)" is suppressed. In each step of the Newton iteration outlined in that section, one has to evaluate the Jacobian matrices $\mathbf{J}^{ij}_{i'j'}$, which are defined by Equation (5.29) for $(i,j) = 11$, 12 and 22, and $(i',j') = 11$, 12 and 22. The elements of these matrices are partial derivatives that according to the chain rule can be expanded as

$$\begin{aligned}\frac{\partial c_{ij,\alpha}}{\partial h_{i'j',\alpha'}} &= \sum_{\beta=0,\beta'=0}^{N_k} \frac{\partial c_{ij,\alpha}}{\partial \tilde{c}_{ij,\beta}} \cdot \frac{\partial \tilde{c}_{ij,\beta}}{\partial \tilde{h}_{i'j',\beta'}} \cdot \frac{\partial \tilde{h}_{i'j',\beta'}}{\partial h_{i'j',\alpha'}} \\ &= \sum_{\beta=0}^{N_k} U_{\alpha\beta} \cdot \frac{\partial \tilde{c}_{ij,\beta}}{\partial \tilde{h}_{i'j',\beta}} \cdot T_{\beta\alpha'} \end{aligned} \tag{C.1}$$

where the second equality follows from the equations

$$\frac{\partial \tilde{c}_{ij,\beta}}{\partial \tilde{h}_{i'j',\beta'}} = \delta_{\beta\beta'} \frac{\partial \tilde{c}_{ij,\beta}}{\partial \tilde{h}_{i'j',\beta}} \tag{C.2}$$

$$T_{\beta'\alpha'} = \frac{\partial \tilde{h}_{i'j',\beta'}}{\partial h_{i'j',\alpha'}} \tag{C.3}$$

$$U_{\alpha\beta} = \frac{\partial c_{ij,\alpha}}{\partial \tilde{c}_{ij,\beta}} \tag{C.4}$$

which are consequences of Equations (5.13), (5.22) and (5.24) respectively. $\delta_{\beta\beta'}$ denotes here the Kronecker delta. In order to obtain the partial derivatives $\partial \tilde{c}_{ij,\beta}/\partial \tilde{h}_{i'j',\beta}$, three linear systems are derived from Equation (5.13) by differentiation with respect to $\tilde{h}_{11,\beta}$, $\tilde{h}_{12,\beta}$ and $\tilde{h}_{22,\beta}$. The obtained systems are

$$(\mathbf{I} + \rho \underline{\mathbf{H}}(\beta \cdot \Delta k)) \begin{pmatrix} \partial \tilde{c}_{11,\beta}/\partial \tilde{h}_{11,\beta} \\ \partial \tilde{c}_{12,\beta}/\partial \tilde{h}_{11,\beta} \\ \partial \tilde{c}_{22,\beta}/\partial \tilde{h}_{11,\beta} \end{pmatrix} = \begin{pmatrix} 1 - x_1 \rho \tilde{c}_{11,\beta} \\ -x_1 \rho \tilde{c}_{12,\beta} \\ 0 \end{pmatrix} \tag{C.5}$$

$$(\mathbf{I} + \rho \underline{\mathbf{H}}(\beta \cdot \Delta k)) \begin{pmatrix} \partial \tilde{c}_{11,\beta}/\partial \tilde{h}_{12,\beta} \\ \partial \tilde{c}_{12,\beta}/\partial \tilde{h}_{12,\beta} \\ \partial \tilde{c}_{22,\beta}/\partial \tilde{h}_{12,\beta} \end{pmatrix} = \begin{pmatrix} -x_2 \rho \tilde{c}_{12,\beta} \\ 1 - x_2 \rho \tilde{c}_{22,\beta} \\ -x_1 \rho \tilde{c}_{12,\beta} \end{pmatrix} \tag{C.6}$$

$$(\mathbf{I} + \rho \underline{\mathbf{H}}(\beta \cdot \Delta k)) \begin{pmatrix} \partial \tilde{c}_{11,\beta}/\partial \tilde{h}_{22,\beta} \\ \partial \tilde{c}_{12,\beta}/\partial \tilde{h}_{22,\beta} \\ \partial \tilde{c}_{22,\beta}/\partial \tilde{h}_{22,\beta} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 - x_2 \rho \tilde{c}_{22,\beta} \end{pmatrix} \tag{C.7}$$

At each iteration step, these three systems are solved for the partial derivatives which then are used to evaluate the Jacobians of Equation (5.29) via Equation (C.1)

D

Angle Averaging of the Intermolecular Potential

For two molecules denoted 1 and 2, being of type i and j, respectively, and with their centers of mass separated by the vector $\boldsymbol{r}_{12}$, the CHARMM force field defines the pair interaction potential as the sum of the Lennard-Jones (LJ) and Coulomb

$$u_{ij}(\boldsymbol{r}_{12}\omega_1\omega_2) = u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) + u_{ij}^{(\mathrm{C})}(\boldsymbol{r}_{12}\omega_1\omega_2) \tag{D.1}$$

where the LJ part is given by

$$u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) = \sum_{\alpha\in M_i,\beta\in M_j} \epsilon_{\alpha\beta}\left[\left(\frac{R_{\min,\alpha\beta}}{|\boldsymbol{r}_{12}-\boldsymbol{r}_{1,\alpha}+\boldsymbol{r}_{2,\beta}|}\right)^{12} -2\left(\frac{R_{\min,\alpha\beta}}{|\boldsymbol{r}_{12}-\boldsymbol{r}_{1,\alpha}+\boldsymbol{r}_{2,\beta}|}\right)^{6}\right] \tag{D.2}$$

where M_i and M_j denote the sets of atoms of molecules of type i and j, respectively, and $\epsilon_{\alpha\beta}$ and $R_{\min,\alpha\beta}$ are CHARMM LJ parameters for interactions between atoms α and β. $\boldsymbol{r}_{1,\alpha}$ denotes a vector pointing from the center of mass of molecule 1 to the location of atom α of the same molecule. $\boldsymbol{r}_{2,\beta}$ is defined likewise for atom β of molecule 2. When r_{12} is large, the first term in the sum is $O(r_{12}^{-12})$ and therefore neglected here. If the denominator of the remaining term is rewritten, one obtains

$$\begin{aligned} &u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) = \\ &-2\sum_{\alpha\in M_i,\beta\in M_j} \frac{\epsilon_{\alpha\beta}R_{\min,\alpha\beta}^6}{\left(r_{12}^2-\boldsymbol{r}_{12}\cdot(\boldsymbol{r}_{2,\beta}-\boldsymbol{r}_{1,\alpha})+|\boldsymbol{r}_{2,\beta}-\boldsymbol{r}_{1,\alpha}|^2\right)^3} + O(r_{12}^{-12}) = \\ &-2\sum_{\alpha\in M_i,\beta\in M_j} \frac{\epsilon_{\alpha\beta}R_{\min,\alpha\beta}^6}{r_{12}^6}\left(1+\frac{\boldsymbol{r}_{12}\cdot(\boldsymbol{r}_{2,\beta}-\boldsymbol{r}_{1,\alpha})+|\boldsymbol{r}_{2,\beta}-\boldsymbol{r}_{1,\alpha}|^2}{r_{12}^2}\right)^{-3} \\ &+O(r_{12}^{-12}) \end{aligned} \tag{D.3}$$

The non-constant term within the parenthesis is $O(r_{12}^{-1})$. Taylor expanding the power function retaining only terms that are $O(r_{12}^{-1})$ leads to

$$\begin{aligned} &u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2) = \\ &-2\sum_{\alpha\in M_i,\beta\in M_j} \frac{\epsilon_{\alpha\beta}R_{\min,\alpha\beta}^6}{r_{12}^6}\left(1-3\frac{\boldsymbol{r}_{12}\cdot(\boldsymbol{r}_{2,\beta}-\boldsymbol{r}_{1,\alpha})}{r_{12}^2}\right) \\ &+O\left(\frac{\max\left(r_{1,\alpha}^2,r_{2,\beta}^2\right)}{r_{12}^8}\right) \end{aligned} \tag{D.4}$$

When the dependence on ω_1 and ω_2 is averaged out, $\boldsymbol{r}_{1,\alpha}$ and $\boldsymbol{r}_{2,\beta}$ vanish for all α and β. Since $\boldsymbol{r}_{12}$ is independent of the orientations, the second term within the parenthesis vanishes as well, and one obtains

$$\left\langle u_{ij}^{(\mathrm{LJ})}(\boldsymbol{r}_{12}\omega_1\omega_2)\right\rangle_{\omega_1\omega 2} = -2\sum_{\alpha\in M_i,\beta\in M_j}\frac{\epsilon_{\alpha\beta}R_{\mathrm{min},\alpha\beta}^6}{r_{12}^6}+O\left(\frac{\max\left(r_{1,\alpha}^2,r_{2,\beta}^2\right)}{r_{12}^8}\right) \tag{D.5}$$

which proves the first half of Equation (5.44). The exact form of the neglected $O(r_{12}^{-8})$ term, as well as terms of higher order, can be evaluated using a procedure given by Gray and Gubbins (1984).

When r_{12} is large and the two molecules both have zero ionic strength, the Coloumbic term is to leading order identical to the dipole-dipole interaction, given by

$$u_{ij}^{(\mathrm{dd})}\left(\boldsymbol{r}_{12}\omega_1\omega_2\right)) = \frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{r_{12}^3}-3\frac{(\boldsymbol{\mu}_1\cdot\boldsymbol{r}_{12})(\boldsymbol{\mu}_2\cdot\boldsymbol{r}_{12})}{r_{12}^5} \tag{D.6}$$

Expressing $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{r}_{12}$ in spherical coordinates (r, θ, ϕ), with the z-axis chosen to lie along the direction of $\boldsymbol{r}_{12}$ (i.e. the intermolecular frame representation of Section 5.1) and simplifying the resulting trigonometric expression leads to

$$u_{ij}^{(\mathrm{dd})}\left(\boldsymbol{r}_{12}\omega_1\omega_2\right) = \frac{\mu_1\mu_2}{r_{12}^3}\left[\cos\left(\phi_2-\phi_1\right)\sin\theta_1\sin\theta_2-2\cos\theta_1\cos\theta_2\right] \tag{D.7}$$

This gives further

$$\left(u_{ij}^{(\mathrm{dd})}\left(\boldsymbol{r}_{12}\omega_1\omega_2\right)\right)^2 = \frac{\mu_1^2\mu_2^2}{r_{12}^6}\left[\cos^2\left(\phi_2-\phi_1\right)\sin^2\theta_1\sin^2\theta_2 +4\cos^2\theta_1\cos^2\theta_2-4\cos\left(\phi_2-\phi_1\right)\sin\theta_1\sin\theta_2\cos\theta_1\cos\theta_2\right] \tag{D.8}$$

When averaging out the dependence on orientations, θ_1, θ_2, ϕ_1 and ϕ_1 are integrated out according to

$$\frac{1}{2}\int_{-1}^{1}d\left(\cos(\theta_1)\right)\frac{1}{2}\int_{-1}^{1}d\left(\cos(\theta_2)\right)\frac{1}{2\pi}\int_0^{2\pi}d\phi_1\frac{1}{2\pi}\int_0^{2\pi}d\phi_2 \tag{D.9}$$

The third term of Equation (D.8) obviously cancels when the ϕ-dependence is integrated out. The other two terms can be determined considering the elementary trigonometric integrals

$$\frac{1}{2}\int_{-1}^{1}d\left(\cos(\theta)\right)\cos^2(\theta) = \frac{1}{3} \tag{D.10}$$

$$\frac{1}{2}\int_{-1}^{1}d\left(\cos(\theta)\right)\sin^2(\theta) = \frac{1}{2}\int_{-1}^{1}d\left(\cos(\theta)\right)\left(1-\cos^2(\theta)\right) = \frac{2}{3} \tag{D.11}$$

$$\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}d\phi_1 d\phi_2\cos^2(\phi_2-\phi_1) = \frac{1}{2} \tag{D.12}$$

Using these identities, Equation (D.8) becomes

$$\left\langle\left(u_{ij}^{(\mathrm{dd})}\left(\boldsymbol{r}_{12}\omega_1\omega_2\right)\right)^2\right\rangle_{\omega_1\omega_2} = \frac{\mu_1^2\mu_2^2}{r_{12}^6}\left[\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{2}{3}+4\cdot\frac{1}{3}\cdot\frac{1}{3}\right] = \frac{2\mu_1^2\mu_2^2}{3r_{12}^6} \tag{D.13}$$

if the orientations are integrated out. This proves Equation (5.36), which is used in the derivation of the DCF tail approximation, Equation (5.44), in Section 5.4

E

Derivations of Thermodynamic Relations

Equation (6.7) Consider for a binary mixture the differential relation for the total volume, V (Smith *et al.*, 2005)

$$dV = \beta V dT - \kappa_T V dP + \bar{v}_1 dN_1 + \bar{v}_2 dN_2 \tag{E.1}$$

where β, T, κ_T, P, $\bar{v}_1$ and N_1 respectively denote thermal expansivity, temperature, pressure, molecular volume of component 1 and the number of molecules of the same component. dV, dT and dN_2 are here set to zero, which gives

$$\left(\frac{\partial P}{\partial N_1}\right)_{N_2,V,T} = \frac{\bar{v}_1}{\kappa_T V} \tag{E.2}$$

The partial derivatives with respect to molecule numbers can according to the chain rule be written as

$$\left(\frac{\partial}{\partial N_1}\right)_{T,V,N_2} = \frac{x_2}{N}\left(\frac{\partial}{\partial x_1}\right)_{T,V,N} + \left(\frac{\partial}{\partial N}\right)_{T,V,x_1} \tag{E.3}$$

$$\left(\frac{\partial}{\partial N_2}\right)_{T,V,N_1} = -\frac{x_1}{N}\left(\frac{\partial}{\partial x_1}\right)_{T,V,N} + \left(\frac{\partial}{\partial N}\right)_{T,V,x_2} \tag{E.4}$$

with $N \equiv N_1 + N_2$ denoting the total number of molecules and x_1 denoting the fraction of molecules of component 1. Inserting these into Equation (E.2) and multiplying by N yields

$$\frac{\rho\bar{v}_1}{\kappa_T} = x_2\left(\frac{\partial P}{\partial x_1}\right)_{N,V,T} + N\left(\frac{\partial P}{\partial N}\right)_{x_1,V,T} = x_2\left(\frac{\partial P}{\partial x_1}\right)_{N,V,T} + \kappa_T^{-1} \tag{E.5}$$

If $\bar{v}_1$ and κ_T are replaced by their DCFI expressions O'Connell (1971b), one obtains Equation (6.7).

Equation (6.6) The DCFIs are related to the molecule number derivatives of the chemical potentials according to O'Connell (1971b)

$$\begin{aligned} k_{\mathrm{B}}T\left(\frac{N}{N_i}\delta_{ij} - C_{ij}\right) &= N\left(\frac{\partial \mu_i}{\partial N_j}\right) = N\left(\frac{\partial^2(NA)}{\partial N_i \partial N_j}\right) = \\ &= N^2\left(\frac{\partial^2 A}{\partial N_i \partial N_j}\right) + N\left(\frac{\partial A}{\partial N_i}\right) + N\left(\frac{\partial A}{\partial N_j}\right) \end{aligned} \tag{E.6}$$

where A and k_B denote the Helmholtz energy per molecule and the Boltzmann constant, respectively, and where all partial derivatives are taken at fixed T, V and remaining molecule numbers. Consider a two-component mixture and evaluate ΔC as defined in Equation (6.6), to obtain

$$\begin{aligned} -\Delta C &= -\left(C_{11} + C_{22} - 2C_{12}\right) = && \text{(E.7)} \\ &= -\frac{N}{N_1} - \frac{N}{N_2} + N^2 \left(\frac{\partial^2}{\partial N_1^2} + \frac{\partial^2}{\partial N_2^2} - 2\frac{\partial^2}{\partial N_1 \partial N_2} \right) A/k_\mathrm{B}T = && \text{(E.8)} \\ &= -\frac{1}{x_1} - \frac{1}{x_2} + N^2 \left(\frac{\partial}{\partial N_1} - \frac{\partial}{\partial N_2} \right)^2 A/k_\mathrm{B}T && \text{(E.9)} \end{aligned}$$

From Equations (E.3) and (E.4), it follows that

$$\left(\frac{\partial}{\partial N_1}\right)_{T,V,N_2} - \left(\frac{\partial}{\partial N_2}\right)_{T,V,N_1} = \frac{1}{N}\left(\frac{\partial}{\partial x_1}\right)_{T,V,N} \tag{E.10}$$

which inserted in Equation (E.7) yields

$$-\Delta C = -\frac{1}{x_1} - \frac{1}{x_2} + \left(\frac{\partial^2 A/k_\mathrm{B}T}{\partial x_1^2}\right)_{T,V,N} \tag{E.11}$$

The molecular Helmholtz energy can as usual be split into excess and ideal-solution parts, $A = A^\mathrm{E} + A^\mathrm{IS}$, where

$$A^\mathrm{IS}(T,V,x_1) = x_1 A_1(T,V) + x_2 A_2(T,V) + k_\mathrm{B}T\left(x_1 \ln x_1 + x_2 \ln x_2\right) \tag{E.12}$$

with A_1 and A_2 denoting the molecular Helmholtz energies of the corresponding pure systems. The 2$^\mathrm{nd}$ derivative of A^IS with respect to x_1 precisely cancels the first two terms on the right-hand side of Equation (E.11), and Equation (6.6) follows.

Equation (6.10) The DCFIs can be expressed in terms of κ_T, v_i and $\left(\frac{\partial \ln \gamma_i}{\partial x_j}\right)_{T,P,N_{k,k\neq j}}$ according to Wooley and O'Connell (1991)

$$C_{11} = 1 - \frac{\rho \bar{v}_1^2}{\kappa_T k_\mathrm{B} T} - x_2 \left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P,N_2} \tag{E.13}$$

$$C_{12} = 1 - \frac{\rho \bar{v}_1 \bar{v}_2}{\kappa_T k_\mathrm{B} T} + x_1 \left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P,N_2} \tag{E.14}$$

$$C_{22} = 1 - \frac{\rho \bar{v}_2^2}{\kappa_T k_\mathrm{B} T} - x_1 \left(\frac{\partial \ln \gamma_2}{\partial x_2}\right)_{T,P,N_1} \tag{E.15}$$

Using these equations to express ΔC, employing the Gibbs-Duhem equation for the activity coefficient derivative, it is found that

$$\Delta C = -\frac{\rho(\bar{v}_1 - \bar{v}_2)^2}{\kappa_T k_\mathrm{B} T} - \frac{1}{x_2}\left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P} \tag{E.16}$$

which in combination with Equation (6.6) proves Equation (6.10).

F

RMSD Plots for CALB Study II

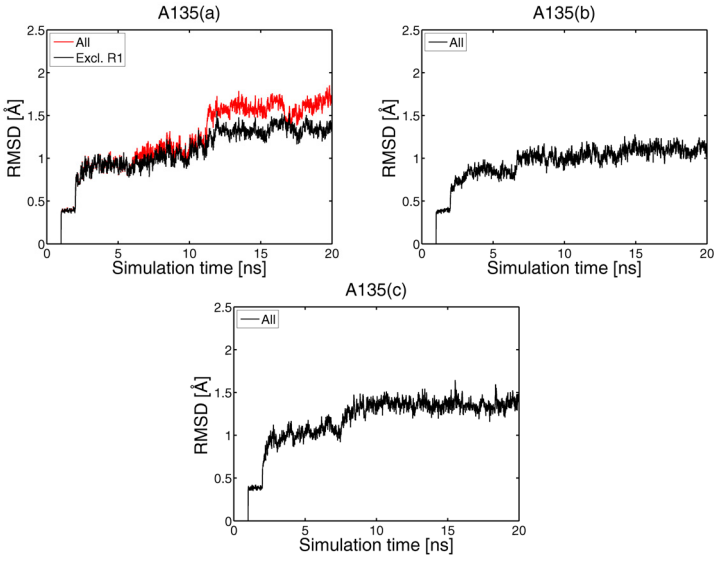


Figure F.1: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 7.1. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 23–32 (R1), 67–75 (R2), 138–152 (R3), 243–292 (R4). N- and C-terminals (1–20 and 308–317) are excluded in all plots.

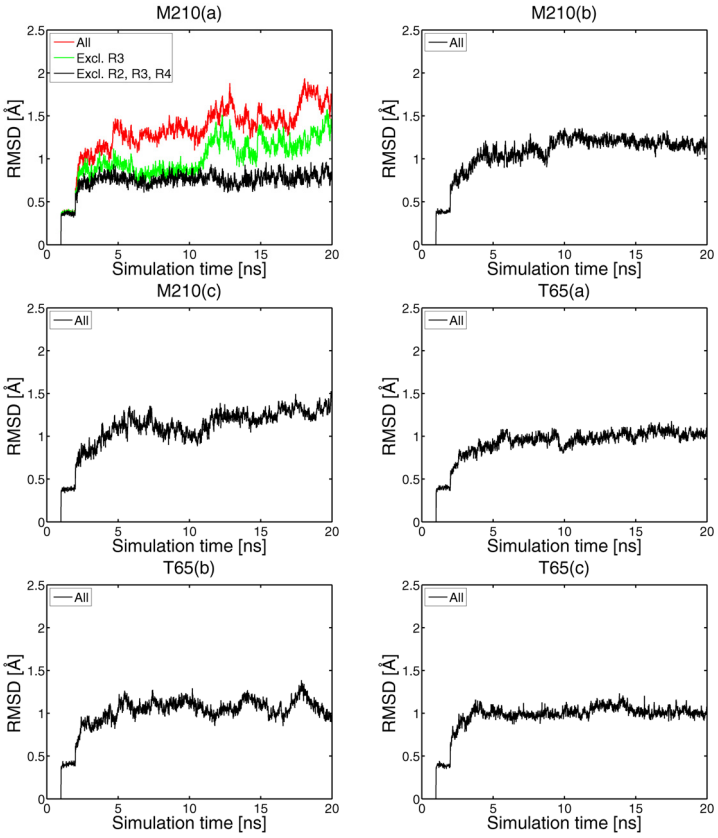


Figure F.2: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 7.1. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 23–32 (R1), 67–75 (R2), 138–152 (R3), 243–292 (R4). N- and C-terminals (1–20 and 308–317) are excluded in all plots.

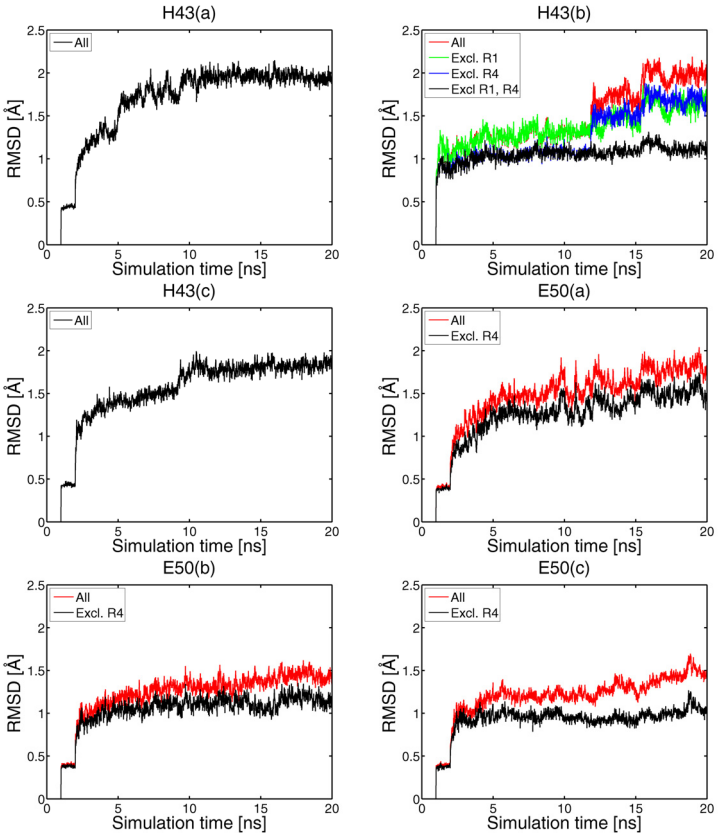


Figure F.3: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 7.1. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 23–32 (R1), 67–75 (R2), 138–152 (R3), 243–292 (R4). N- and C-terminals (1–20 and 308–317) are excluded in all plots.

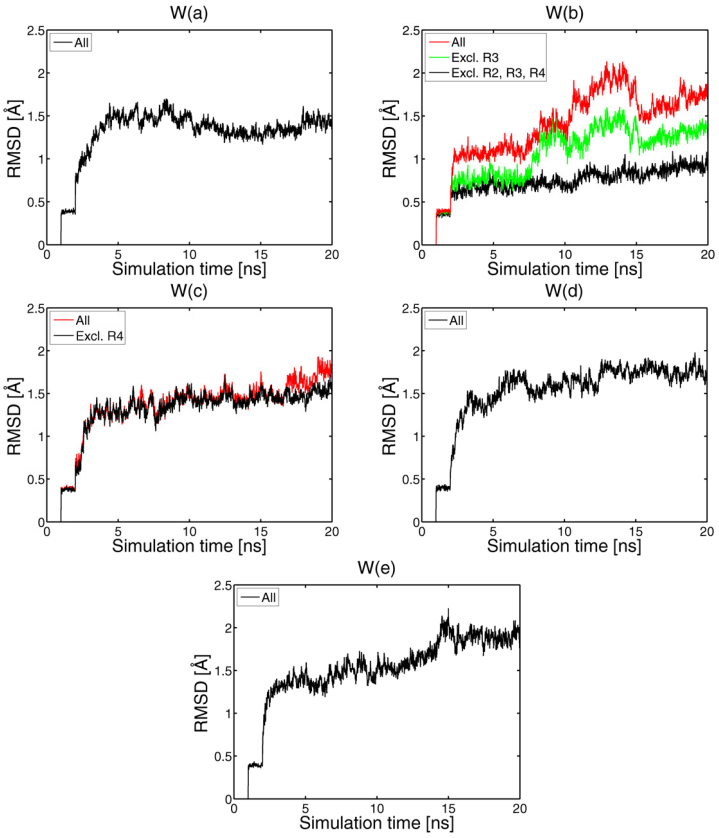


Figure F.4: RMSD plots for selected simulations. Figure titles give the simulation identifiers defined in Table 7.1. Regions that need to be excluded from the calculation for obtaining a stable RMSD, in one or several simulations, comprise residues 23–32 (R1), 67–75 (R2), 138–152 (R3), 243–292 (R4). N- and C-terminals (1–20 and 308–317) are excluded in all plots.

References

Abbott, M. M. and van Ness, H. C. (1975). Vapor–liquid equilibrium: Part iii. Data reduction with precise expressions for G^E. *AIChE J.*, **21**, 62–71.

Abildskov, J., Ellegaard, M. D., and O'Connell, J. P. (2009). Correlation of phase equilibria and liquid densities for gases with ionic liquids. *Fluid Phase Equilib.*, **286**, 95–106.

Abildskov, J., Ellegaard, M. D., and O'Connell, J. P. (2010a). Densities and isothermal compressibilities of ionic liquids – Modeling and application. *Fluid Phase Equilib.*, **295**, 215–229.

Abildskov, J., Wedberg, R., Peters, G. H., van Leeuwen, Boeriu, C. G., and van den Broek (2010b). Solvent selection for biocatalysis. A multiscale methodolgy. *In preparation*.

Achenie, L. E. K., Gani, R., and Venkatasubramanian, V. (2003). *Computer aided molecular design: Theory and practice*. Elsevier.

Affleck, R., Xu, Z., Suzawa, V., Focht, K., Clarck, D. S., and Dordick, J. S. (1992). Enzymatic catalysis and dynamics in low-water environments. *PNAS*, **89**, 1100–1104.

Akoh, C., Chang, S., Lee, G., and Shaw, J. (2007). Enzymatic approach to biodiesel production. *J. Agric. Food Chem.*, **55**, 8995–9005.

Allen, M. P. and Tildesley, D. J. (1987). *Computer Simulation of Liquids*. Oxford University Press, New York.

Anderson, E. M., Larsson, K. M., and Kirk, O. (1998). One biocatalyst - many applications: The use of *CANDIDA ANTARCTICA* B-lipase in organic synthesis. *Biocatal. Biotransform.*, **16**, 181–204.

Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008). Very fast prediction and rationalization of pK_a values for protein-ligand complexes. *Proteins*, **73**, 765–783.

Bell, G., Janssen, A. E. M., and Halling, P. J. (1997). Water activity fails to predict critical hydration level for enzyme activity in polar organic solvents: Interconversion of water concentrations and activities. *Enzyme Microb. Technol.*, **20**, 471–477.

Bellot, J. C., Choisnard, L., Castillo, E., and Marty, A. (2001). Combining solvent engineering and thermodynamic modeling to enhance selectivity during monoglyceride synthesis by lipase-catalyzed esterification. *Enzyme Microb. Technol.*, **28**, 362–369.

Ben-Naim, A. (2008). The Kirkwood-Buff integrals of one-component liquids. *J. Chem. Phys.*, **23**, 234501.

Berger, M. and Schneider, M. (1992). Enzymatic esterification of glycerol ii. lipase-catalyzed synthesis of regioisomerically pure l(3)-rac-monoacylglycerols. *J. Am. Oil Chem. Soc.*, **69**, 961–965.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bovara, R., Carrea, G., Ottolina, G., and Riva, S. (1993). Effects of water activity on v_{max}, and k_m, of lipase-catalyzed transesterification in organic media. *Biotechnol. Lett.*, **15**, 937–942.

Branco, R. J. F., Graber, M., Denis, V., and Pleiss, J. (2009). Molecular mechanism of the hydration of *Candida antarctica* lipase B in the gas phase: Water adsorption isotherms and molecular dynamics simulations. *ChemBioChem*, **10**, 2913–2919.

Brändén, C. and Tooze, J. (1999). *Introduction to protein structure*. Garland Pub.

Brelvi, S. W. and O'Connell, J. P. (1972). A corresponding states correlation of liquid compressibilities and partial molar volumes of gases at infinite dilution. *AIChE J.*, **18**, 1239–1243.

Broos, J., Visser, A. J. W. G., Engbersen, J. F. J., Verboom, W., van Hoek, A., and Reinhoudt, D. N. (1995). Flexibility of enzymes suspended in organic solvents probed by time-resolved fluorescence anisotropy. evidence that enzyme activity and enantioselectivity are directly related to enzyme flexibility. *J. Am. Chem. Soc.*, **117**, 12657–12663.

Butler, R. M., Cooke, G. M., Lukk, G. G., and Jameson, B. G. (1956). Prediction of flash points of middle distillates. *Ind. Eng. Chem.*, **48**, 808–812.

Carrea, G. and Riva, S. (2000). Properties and synthetic applications of enzymes in organic solvents. *Angew. Chem. Int. Ed.*, **39**, 2226–2254.

Carrea, G., Ottolina, G., and Riva, S. (1995). Role of solvents in the control of enzyme selectivity in organic media. *Trends Biotechnol.*, **13**, 63–70.

Chamouleau, F., Coulon, D., Girardin, M., and Ghoul, M. (2001). Influence of water activity and water content on sugar esters lipase-catalyzed synthesis in organic media. *J. Mol. Cat. B: Enzym.*, **11**, 949–954.

Cherukuvada, S. L., Seshasayee, A. S. N., Raghunathan, K., Anishetty, S., and Pennathur, G. (2005). Evidence of a double-lid movement in pseudomonas aeruginosa lipase: Insights from molecular dynamics simulations. *PLoS Comp. Biol.*, **1**, 0182–0189.

Christensen, S., Peters, G. H., Hansen, F. Y., O'Connell, J. P., and Abildskov, J. (2007a). Generation of thermodynamics data for organic liquid mixtures from molecular simulations. *Mol. Simul.*, **33**, 449–457.

Christensen, S., Peters, G. H., Hansen, F. Y., O'Connell, J. P., and Abildskov, J. (2007b). State conditions transferability of vapor-liquid equilibria via fluctuation solution theory with correlation function integrals from molecular dynamics simulation. *Fluid Phase Equilib.*, **260**, 169–176.

Christensen, S., Peters, G. H., Hansen, F. Y., and Abildskov, J. (2007c). Thermodynamic models from fluctuation solution theory analysis of molecular simulations. *Fluid Phase Equilib.*, **261**, 185–190.

Colombo, G., Ottolina, G., Carrea, G., Bernardi, A., and C., S. (1998). Application of structure-based thermodynamic calculations to the rationalization of the enantioselectivity of subtilisin in organic solvents. *Tetrahedron: Asymmetry*, **9**, 1205–1214.

Colombo, G., Toba, S., and Merz, K. M. (1999). Rationalization of the enantioselectivity of Subtilisin in DMF. *J. Am. Chem. Soc.*, **121**, 3486–3493.

Colombo, G., Ottolina, G., Carrea, G., and Merz, K. M. (2000). Modelling the enantioselectivity of subtilisin in water and organic solvents: Insights from molecular dynamics and quantum mechanical/molecular mechanical studies. *Chem. Commun.*, pages 559–560.

Constantinou, L. and Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE J.*, **40**, 1697–1710.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.

Cruz, A., Ramirez, E., Santana, A., Barletta, G., and López, G. E. (2009). Molecular dynamic study of subtilisin carlsberg in aqueous and nonaqueous solvents. *Mol. Simul.*, **35**, 205–212.

Damstrup, M., Abildskov, J., Kiil, S., Jensen, A. D., Sparsø, F. V., and Xu, X. (2006). Evaluation of binary solvent mixtures for efficient monoacylglycerol production by continuous enzymatic glycerolysis. *J. Agric. Food Chem.*, **54**, 7113–7119.

Damstrup, M. L., Jensen, T., Sparsø, F. V., Kiil, S., Jensen, A. D., and Xu, X. (2005). Solvent optimization for efficient enzymatic monoacylglycerol production based on a glycerolysis reaction. *J. Am. Oil Chem. Soc.*, **82**, 559–564.

de Leeuw, S. W., Smit, B., and Williams, C. P. (1990). Molecular dynamics studies of polar/nonpolar fluid mixtures. I. Mixtures of Lennard-Jones and Stockmayer fluids. *J. Chem. Phys.*, **93**, 2704–2714.

Degn, P. and Zimmermann, W. (2001). Optimization of carbohydrate fatty acid ester synthesis in organic media by a lipase from *Candida antarctica*. *Biotechnol. Bioeng.*, **74**, 483–491.

Derewenda, Z. S. (1994). Structure and function of lipases. *Adv. Protein Chem.*, **45**, 1–52.

Díaz-Vergara, N. and Piñeiro, Á. (2008). Molecular dynamics study of triosephosphate isomerase from *Trypanosoma cruzi* in water/decane mixtures. *J. Phys. Chem. B*, **112**, 3529–3539.

Dodson, G. G., Lawson, D. M., and Winkler, F. K. (1992). Structural and evolutionary relationships in lipase mechanism and activation. *Faraday Discuss.*, **93**, 95–105.

Dordick, J. S. (1992). Designing enzymes for use in organic solvents. *Biotechnol. Prog.*, **8**, 259–267.

Dordick, J. S., Marletta, M. A., and Klibanov, A. M. (1986). Peroxidases depolymerize lignin in organic media but not in water. *PNAS*, **83**, 6255–6257.

Engel, T. and Hehre, W. (2005). *Quantum Chemistry and Spectroscopy*. Pearson-/Benjamin Cummings.

Fernandez-Ramos, A., Ellingson, B. A., Garrett, B. C., and Truhlar, D. G. (2007). *Reviews in Compuational Chemistry*, volume 23, chapter 3, pages 125–232. Wiley-VCH.

Fersht, A. (1999). *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*. W. H. Freeman.

Fischer, K. and Gmehling, J. (1994). P-x and γ^{∞} data for the different binary butanol-water systems at 50 °C. *J. Chem. Eng. Data*, **39**, 309–315.

Fitzpatrick, P. L., Steinmetz, A. C. U., Ringe, D., and Klibanov, A. M. (1993). Enzyme crystal structure in a neat organic solvent. *PNAS*, **90**, 8653–8657.

Fitzpatrick, P. L., Ringe, D., and Klibanov, A. M. (1994). X-ray crystal-structure of cross-linked subtilisin Carlsberg in water versus acetonitrile. *Biochem. Biophy. Res. Commun.*, **198**, 675–681.

Fjerbaek, L., Christensen, K. V., and Norddahl, B. (2009). A review of the current state of biodiesel production using enzymatic transesterification. *Biotechnol. Bioeng.*, **102**, 1298–1315.

Foresti, M. L., Galle, M., Ferreira, M. L., and Briand, L. E. (2009). Enantioselective esterification of ibuprofen with ethanol as reactant and solvent catalyzed by immobilized lipase: Experimental and molecular modeling aspects. *J. Chem. Technol. Biotechnol.*, **84**, 1461–1473.

Frauenfelder, H. (2008). What determines the speed limit on enzyme catalysis? *Nat. Chem. Biol.*, **4**, 21–22.

Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation*. Calif.

Fries, P. H. and Patey, G. N. (1985). The solution of the hypernetted-chain approximation for fluids of nonspherical particles. A general method with application to dipolar hard spheres. *J. Chem. Phys.*, **82**, 429–440.

Gao, J. and Truhlar, D. G. (2002). Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.*, **53**, 467505.

Goldstein, H. (1980). *Classical Mechanics*. Addison-Wesley.

Graber, M., Bousquet-Dubouch, M., Lamare, S., and Legoy, M. (2003). Alcoholysis catalyzed by *Candida antarctica* lipase B in a gas/solid system: Effects of water on kinetic parameters. *Biochim. Biophys. Acta*, **1648**, 24–32.

Graber, M., Irague, R., Rosenfeld, E., Lamare, S., Franson, L., and Hult, K. (2007). Solvent as a competitive inhibitor for *Candida antarctica* lipase B. *Biochim. Biophys. Acta*, **1774**, 1052–1057.

Gray, G. and Gubbins, K. E. (1984). *Theory of Molecular Fluids. Volume I: Fundamentals*. Oxford University Press, New York.

Griebenow, K. and Klibanov, A. M. (1996). On protein denaturation in aqueous-organic mixtures but not in pure organic solvents. *J. Am. Chem. Soc.*, **118**, 11695–11700.

Gross, J. and Vrabec, J. (2006). An equation of state contribution for polar components: Dipolar molecules. *AIChE J.*, **52**, 1194–1204.

Gubbins, K. E. and O'Connell, J. P. (1974). Isothermal compressibility and partial molal volume for polyatomic liquids. *J. Chem. Phys.*, **60**, 3449–3453.

Halling, P. J. (1989). Organic liquids and biocatalysts: Theory and practice. *Trends Biotechnol.*, **7**, 50–52.

Halling, P. J. (1990a). High-affinity binding of water by proteins is similar in air and in organic solvents. *Biochim. Biophys. Acta*, **1040**, 225–228.

Halling, P. J. (1990b). Solvent selection for biocatalysis in mainly organic systems: Prediction of effects on equilibrium position. *Biotechnol. Bioeng.*, **35**, 691–70.

Halling, P. J. (1994). Thermodynamic predictions for biocatalysis in nonconvetional media: Theory, tests, and recommendations for experimental design and analysis. *Enzyme Microb. Technol.*, **16**, 178–206.

Handa, Y. P. and Benson, G. C. (1979). Volume changes on mixing two liquids: A review of the experimental techniques and the literature data. *Fluid Phase Equilib.*, **3**, 185–249.

Hansen, H. K., Rasmussen, P., Fredenslund, A., Schiller, M., and Gmehling, J. (1991). Vapor-liquid equilibria by unifac group contribution. 5. Revision and extension. *Ind. Eng. Chem. Res.*, **30**, 2352–2355.

Hartsough, D. S. and Merz, K. M. (1992). Protein flexibility in aqueous and non-aqueous solutions. *J. Am. Chem. Soc.*, **114**, 10113–10116.

Hartsough, D. S. and Merz, K. M. (1993). Protein dynamics and solvation in aqueous and nonaqueous environments. *J. Am. Chem. Soc.*, **115**, 6529–6537.

Hess, B. and van der Vegt, N. F. A. (2009). Cation specific binding with protein surface charges. *PNAS*, **106**, 13296–13300.

Hu, J., Ma, A., and Dinner, A. R. (2006). Monte carlo simulations of biomolecules: The MC module in CHARMM. *J. Comput. Chem.*, **27**, 203–216.

Huang, Y. H. (1986). *Thermodynamic properties of compressed liquids and liquid mixtures from fluctuation solution theory*. Ph.D. thesis, University of Florida.

Huang, Y. H. and O'Connell, J. P. (1987). Corresponding states correlation for the volumetric properties of compressed liquids and liquid mixtures. *Fluid Phase Equilib.*, **37**, 75–84.

Humeau, C., Girardin, M., Rovel, B., and Miclo, A. (1998). Effect of the thermodynamic water activity and the reaction medium hydrophobicity on the enzymatic synthesis of ascorbyl palmitate. *J. Biotechnol.*, **63**, 1–8.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *J. Molec. Graphics*, **14**, 33–38.

Jääskeläinen, S., Verma, C. S., Hubbard, R. E., Linko, P., and Caves, L. S. D. (1998). Conformational change in the activation of lipase: An analysis in terms of low-frequency normal modes. *Prot. Sci.*, **7**, 1359–1367.

James, J. J., Lakshmi, B. S., Seshasayee, A. S. N., and Gautam, P. (2007). Activation of candida rugosa lipase at alkaneaqueous interfaces: A molecular dynamics study. *FEBS Lett.*, **581**, 4377–4383.

Ji, J., Çağin, T., and Pettitt, B. M. (1992). Dynamic simulations of water at constant chemical potential. *J. Chem. Phys.*, **96**, 1333–1342.

Jolly, D. L., Freasier, B. C., and Bearman, R. J. (1976). The extension of simulation radial distribution functions to an arbitrary range by baxter's factorisation technique. *Chem. Phys.*, **15**, 237–242.

Kaewthong, W. and H-Kittikun, A. (2004). Glycerolysis of palm olein by immobilized lipase PS in organic solvents. *Enzyme Microb. Technol.*, **35**, 218–222.

Kaieda, M., Samukawa, T., Kondo, A., and Fukuda, H. (2001). Effect of methanol and water contents on production of biodiesel fuel from plant oil catalyzed by various lipases in a solvent-free system. *J. Biosci. Bioeng.*, **91**, 12–15.

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, **105**, 6474–6487.

Kazandjian, R. Z., Dordick, J. S., and Klibanov, A. M. (1986). Enzymatic analyses in organic solvents. *Biotechnol. Bioeng.*, **28**, 417–421.

Ke, T., Wescott, C. R., and Klibanov, A. M. (1996). Prediction of the solvent dependence of enzymatic prochiral selectivity by means of structure-based thermodynamic calculations. *J. Am. Chem. Soc.*, **118**, 3366–3374.

Kim, J., Clarck, D. S., and Dordick, J. S. (2000). Intrinsic effects of solvent polarity on enzymic activation energies. *Biotechnol. Bioeng.*, **67**.

Kirkwood, J. G. and Buff, F. P. (1951). The statistical mechanical theory of solutions. I. *J. Chem. Phys.*, **19**, 774–777.

Klibanov, A. M. (1997). Why are enzymes less active in organic solvents than in water? *Trends Biotechnol.*, **15**, 97–101.

Klibanov, A. M. (2001). Improving enzymes by using them in organic solvents. *Nature*, **409**, 241–246.

Kurihara, K., Minoura, T., Takeda, K., and Kojima, K. (1995). Isothermal vapor-liquid-equilibria for methanol plus ethanol plus water, methanol plus water, and ethanol plus water. *J. Chem. Eng. Data*, **40**, 679–684.

Laane, C., Boeren, S., Vos, K., and Veeger, C. (1987). Rules for optimization of biocatalysis in organic solvents. *Biotechnol. Bioeng.*, **30**, 81–87.

Laidler, K. J. (1987). *Chemical Kinetics*. Harper and Row, New York.

Lebowitz, J. L. and Percus, J. K. (1963a). Asymptotic behavior of the radial distribution functions. *J. Math. Phys.*, **4**, 248–254.

Lebowitz, J. L. and Percus, J. K. (1963b). Statistical thermodynamics of nonuniform fluids. *J. Math. Phys.*, **4**, 116–123.

Leonard-Nevers, V., Martona, Z., Lamarea, S., Hult, K., and Graber, M. (2009). Understanding water effect on *Candida antarctica* lipase B activity and enantioselectivity towards secondary alcohols. *J. Mol. Cat. B: Enzym.*, **59**, 90–95.

Li, H., Robertson, A. D., and Jensen, J. H. (2005). Very fast empirical prediction and interpretation of protein pK_a values. *Proteins*, **61**, 704–721.

Lynch, G. C. and Pettitt, B. M. (1997). Grand canonical ensemble molecular dynamics simulations: Reformulation of extended system dynamics approaches. *J. Chem. Phys.*, **107**, 8594–8610.

MacKerell Jr., A. D., Bashford, D., Bellot Jr., M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnik, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher III, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wio'rkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics of proteins. *J. Phys. Chem.*, **102**, 3586–3616.

MacKerell Jr., A. D., Feig, M., and Brooks III, C. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, **25**, 1400–1415.

Maginn, E. (2009). From discovery to data: What must happen for molecular simulation to become a mainstream chemical engineering tool. *AIChE J.*, **55**, 1304–1310.

Magnusson, A. O., Rotticci-Mulder, J. C., Santagostino, A., and Hult, K. (2005). Creating space for large secondary alcohols by rational redesign of *Candida antarctica* lipase B. *ChemBioChem*, **6**, 1051–1056.

Makarov, V. A., Andrews, B. K., Smith, P. E., and Pettitt, B. M. (2000). Residence times of water molecules in the hydration sites of myoglobin. *Biophys. J.*, **79**, 2966–2974.

Martin, M. and Biddy, M. (2005). Monte Carlo molecular simulation predictions for the heat of vaporization of acetone and butyramide. *Fluid Phase Equilib.*, **236**, 53–57.

Martin, M. G. and Siepmann, J. I. (1999). Novel configurational-bias Monte Carlo method for branched molecules. Transferable potentials for phase equilibria. II. United-atom description of branched alkanes. *J. Phys. Chem. B*, **103**, 4508–4517.

Martin, T. M. and Young, D. M. (2001). Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (*Pimephales promelas*) using a group contribution method. *Chem. Res. Toxicol.*, **14**, 1378–1385.

Martinelle, M. and Hult, K. (1995). Kinetics of acyl transfer reaction in organic media catalysed by *Candida antarctica* lipase B. *Biochim. Biophys. Acta*, **1251**, 191–197.

Martinelle, M., Holmquist, M., and Hult, K. (1995). On the interfacial activation of *Candida antarctica* lipase B. *Biochim. Biophys. Acta*, **1258**, 272–276.

Matteoli, E. and Mansoori, G. A. (1995). A simple expression for radial distribution functions of pure fluids and mixtures. *J. Chem. Phys.*, **103**, 4672–4677.

McCabe, R. W., Rodger, A., and Taylor, A. (2005). A study of the secondary structure of *Candida antarctica* lipase B using synchrotron radiation circular dichroism measurements. *Enzyme Microb. Technol.*, **36**, 70–74.

McQuarrie, D. A. (1976). *Statistical Mechanics*. Harper & Row, New York.

Mecke, M., Müller, A., Winkelmann, J., Vrabec, J., Fischer, J., Span, R., and Wagner, W. (1996). An accurate van der Waals-type equation of state for the Lennard-Jones fluid. *Int. J. Thermohys.*, **17**, 391–404.

Micaêlo, N. M. and Soares, C. M. (2007). Modeling hydration mechanisms of enzymes in nonpolar and polar organic solvents. *FEBS J.*, **274**, 2424–2436.

Micaêlo, N. M. and Soares, C. M. (2008). Protein structure and dynamics in ionic liquids. Insights from molecular dynamics simulation studies. *J. Phys. Chem. B*, **112**, 2566–2572.

Micaêlo, N. M., Teixeira, V. H., Baptista, A. M., and Soares, C. M. (2005). Water dependent properties of cutinase in nonaqueous solvents: A computational study of enantioselectivity. *Biophys. J.*, **89**, 999–1008.

Mikhail, S. Z. and Kimel, W. R. (1961). Densities and viscosities of methanol-water mixtures. *J. Chem. Eng. Data*, **6**, 533–537.

Mora-Pale, J. M., Pérez-Munguía, S., González-Mejía, J. C., Dordick, J. S., and Bárzana, E. (2007). The lipase-catalyzed hydrolysis of lutein diesters in non-aqueous media is favored at extremely low water activities. *Biotechnol. Bioeng.*, **98**, 535–542.

Murad, S., Gubbins, K. E., and Gray, C. G. (1983). Comparisons of perturbation and integral equation theories for the angular pair correlation function in molecular fluids. *Chem. Phys.*, **81**, 87–98.

Nichols, J. W., Moore, S. G., and Wheeler, D. R. (2009). Improved implementation of Kirkwood-Buff solution theory in periodic molecular simulations. *Phys. Rev. E*, **80**, 051203.

Nienhuis, G. and Deutch, J. M. (1971). Structure of dielectric fluids. I. The two-particle distribution function of polar fluids. *J. Chem. Phys.*, **55**, 4213–4229.

Nordblad, M. and Adlercreutz, P. (2008). Effects of acid concentration and solvent choice on enzymatic acrylation by *Candida antarctica* lipase B. *J. Biotechnol.*, **133**, 127–133.

Norin, M., Haeffner, F., Hult, K., and Edholm, O. (1994). Molecular dynamics simulations of an enzyme surrounded by vacuum, water, or a hydrophobic solvent. *Biophys. J.*, **67**, 548–559.

O'Connell, J. P. (1971a). Molecular thermodynamics of gases in mixed solvents. *AIChE J.*, **17**, 658–663.

O'Connell, J. P. (1971b). Thermodynamic properties of solutions based on correlation functions. *Mol. Phys.*, **20**, 27–33.

O'Connell, J. P. (1994). Thermodynamics and fluctuation solution theory with some applications to systems at near- or supercritical conditions. *NATO Adv Sci I E-App*, **273**, 191.

Odele, O. and Macchietto, S. (1993). Computer aided molecular design: A novel method fo optimal solvent selection. *Fluid Phase Equilib.*, **82**, 47–54.

Ornstein, L. S. and Zernike, F. (1914). Accidental deviations of density and opalescence at the critical point of a single substance. *Proc. K. Ned. Akad. Wet.*, **17**, 793–806.

Panagiotopoulos, A. Z. (1987a). Adsorption and capillary condensation of fluids in cylindrical pores by Monte Carlo simulation in the Gibbs ensemble. *Mol. Phys.*, **62**, 701–719.

Panagiotopoulos, A. Z. (1987b). Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble. *Mol. Phys.*, **61**, 813–826.

Panagiotopoulos, A. Z., Quirke, N., Stapleton, M., and Tildesley, D. J. (1988). Phase equilibria by simulation in the Gibbs ensemble. Alternative derivation, generalization and application to mixture and membrane equilibria. *Mol. Phys.*, **63**, 527–545.

Perera, A. and Sokolić, F. (2004). Modeling nonionic solutions: The acetone-water mixture. *J. Chem. Phys.*, **121**, 11272–11282.

Peters, G. H., van Aalten, D. M. F., Edholm, O., Toxværd, S., and Bywater, R. (1996a). Dynamics of proteins in different solvent systems: Analysis of essential motion in lipases. *Biophys. J.*, **71**, 2245–2255.

Peters, G. H., Olsen, O. H., Svendsen, A., and Wade, R. C. (1996b). Theoretical investigation of the dynamics of the active site lid in *Rhizomucor miehei* lipase. *Biophys. J.*, **71**, 119–129.

Peters, G. H., Toxværd, S., Olsen, O. H., and Svendsen, A. (1997). Computational studies of the activation of lipases and the effect of a hydrophobic environment. *Protein Eng.*, **10**, 137–147.

Petersson, A. E. V., Adlercreutz, P., and Mattiasson, B. (2006). A water activity control system for enzymatic reactions in organic media. *Biotechnol. Bioeng.*, **97**, 235–241.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.

Pleiss, J., Fischer, M., and Schmid, R. D. (1998). Anatomy of lipase binding sites: The scissile fatty acid binding. *Chem. Phys. Lipids*, **93**, 67–80.

Poling, B. E., Prausnitz, J. M., and O'Connell, J. P. (2007). *The properties of gases and liquids*. McGraw-Hill.

Ponder, J. W. and Case, D. A. (2003). Force fields for protein simulations. *Adv. Protein Chem.*, **66**, 27–85.

Ramirez, R., Mareschal, M., and Borgis, D. (2005). Direct correlation functions and the density functional theory of polar solvents. *J. Chem. Phys.*, **319**, 261–272.

Rapaport, D. C. (2004). *The Art of Molecular Dynamics Simulation*. Cambridge University Press, Cambridge.

Reed, T. M. and Gubbins, K. E. (1973). *Applied statistical mechanics*. McGraw-Hill, New York.

Rekker, R. F. and de Kort, H. M. (1979). The hydrophobic fragmental constant, an extension to a 1000 datapoint set. *Eur. J. Med. Chem.*, **14**, 479–488.

Rendón, X., López-Munguía, A., and Castillo, E. (2001). Solvent engineering applied to lipase-catalyzed glycerolysis of triolein. *J. Am. Oil Chem. Soc.*, **78**, 1061–1066.

Roccatano, D. (2008). Computer simulations study of biomolecules in non-aqueous or cosolvent/water mixture solutions. *Curr. Protein. Pept. Sci.*, **9**, 407–426.

Rod, T. H. and Ryde, U. (2005a). Accurate QM/MM free energy calculations of enzyme reactions: Methylation by Catechol O-Methyltransferase. *J. Chem. Theory Comput.*, **1**, 1240–1251.

Rod, T. H. and Ryde, U. (2005b). Quantum mechanical free energy barrier for an enzymatic reaction. *Phys. Rev. Lett.*, **94**, 138302.

Rossky, P. J. (1985). The structure of polar molecular liquids. *Annu. Rev. Phys. Chem.*, **36**, 321–346.

Rowlinson, J. S. (1965). The equation of state of dense systems. *Rep. Prog. Phys.*, **28**, 169–199.

Salacuse, J. J., Denton, A. R., and Egelstaff, P. A. (1996). Finite-size effects in molecular dynamics simulations: Static structure factor and compressibility. I. Theoretical method. *Phys. Rev. E*, **53**, 2382–2389.

Schröder, C., Rudas, T., Boresch, S., and Steinhauser, O. (2006). Simulation studies of the protein-water interface. I. Properties at the molecular resolution. *J. Chem. Phys.*, **124**, 234907.

Senn, H. M. and Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.*, **48**, 1198–1229.

Skjøt, M., De Maria, L., Chatterjee, R., Svendsen, A., Patkar, S. A., Østergaard, P. R., and Brask, J. (2009). Understanding the plasticity of the α/β hydrolase fold: Lid swapping on the *Candida antarctica* lipase B. Results in chimeras with interesting biocatalytic properties. *ChemBioChem*, **10**, 520–527.

Smit, B. (1992). Phase diagrams of Lennard-Jones fluids. *J. Chem. Phys.*, **96**, 8639–8640.

Smith, J. M., van Ness, H. C., and Abbott, M. M. (2005). *Introduction to Chemical Engineering Thermodynamics*. McGraw-Hill.

Soares, C. M., Teixeira, V. H., and Baptista, A. M. (2003). Protein structure and dynamics in nonaqueous solvents: Insights from molecular dynamics simulation studies. *Biophys. J.*, **84**, 1628–1641.

Stamatis, H., Voutsas, E. C., Delimitsou, C., Kolisis, F. N., and Tassios, D. (2000). Enzymatic production of alkyl esters through lipase-catalyzed transesterification reactions in organic solvents: Solvent effects and prediction capabilities of equilibrium constants. *Biocatal. Biotransform.*, **18**, 259–269.

Su, E. and Wei, D. (2008). Improvement in lipase-catalyzed methanolysis of triacylglycerols for biodiesel production using a solvent engineering method. *J. Mol. Catal. B: Enzym.*, **55**, 118–125.

Taylor, A. E. (1900). Vapour-pressure relations in mixtures of two liquids. *J. Phys. Chem.*, **4**, 290–305.

Tejo, B. A., Salleh, A. B., and Pleiss, J. (2004). Structure and dynamics of *Candida rugosa* lipase: The role of organic solvent. *J. Mol. Model.*, **10**, 358–366.

Toba, S. and Merz, K. M. (1996). Solvation and dynamics of chymotrypsin in hexane. *J. Am. Chem. Soc.*, **118**, 6490–6498.

Toba, S. and Merz, K. M. (1997). The concept of solvent compatibility and its impact on protein stability and activity enhancement in nonaqueous solvents. *J. Am. Chem. Soc.*, **119**, 9939–9948.

Trodler, P. and Pleiss, J. (2008). Modeling structure and flexibility of *Candida antarctica* lipase B in organic solvents. *BMC Struct. Biol.*, **8**.

Trodler, P., Schmid, R. D., and Pleiss, J. (2009). Modeling of solvent-dependent conformational transitions in *Burkholderia cepacia* lipase. *BMC Struct. Biol.*, **9**.

Uppenberg, J., Hansen, M. T., Patkar, S., and Jones, T. A. (1994). The sequence, crystal structure determination and refinement of two crystal forms of lipase B from *Candida antarctica*. *Structure*, **2**, 293–308.

Uppenberg, J., Örhner, N., Norin, M., Hult, K., Kleywegt, G. J., Patkar, S., Waagen, V., Anthonsen, T., and Jones, T. A. (1995). Crystallographic and molecular-modeling studies of lipase B from *Candida Antarctica* reveal a stereospecific pocket for secondary alcohols. *Biochemistry*, **34**, 16838–16851.

Valivety, R. H., Johnston, G. A., Suckling, C. J., and Halling, P. J. (1991). Solvent effects on biocatalysis in organic systems: Equilibrium position and rates of lipase catalyzed esterification. *Biotechnol. Bioeng.*, **38**, 1137–1143.

Valivety, R. H., Halling, P. J., Peilow, A. D., and Macrae, A. R. (1992a). Lipases from different sources vary widely in dependence of catalytic activity on water activity. *Biochim. Biophys. Acta*, **1122**, 143–146.

Valivety, R. H., Halling, P. J., and Macrae, A. R. (1992b). Reaction rate with suspended lipase catalyst shows similar dependence on water activity in different organic solvents. *Biochim. Biophys. Acta*, **1118**, 218–222.

Valivety, R. H., Halling, P. J., and Macrae, A. R. (1993). Water as a competitive inhibitor of lipase-catalysed esterification in organic media. *Biotechnol. Lett.*, **15**, 1133–1138.

van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Makr, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, **26**, 1701–1718.

van Ness, H. C. (1995). Thermodynamics in the treatment of (vapor + liquid) equilibria. *J. Chem. Thermodynamics*, **27**, 113–134.

Verlet, L. (1968). Computer "experiments" on classical fluids. II. Equilibrium correlation functions. *Phys. Rev.*, **165**, 201–214.

Vidinha, P., Harper, N., Micaelo, N. M., Lourenco, N. M. T., Gomes da Silva, M. D. R., Cabral, J. M. S., Afonso, C. A. M., Soares, C. M., and Barreiros, S. (2003). Effect of immobilization support, water activity, and enzyme ionization state on cutinase activity and enantioselectivity in organic media. *Biotechnol. Bioeng.*, **85**, 442–449.

Vorobyov, I., Anisimov, V., Greene, S., Venable, R., Moser, A., Pastor, R., and MacKerell Jr., A. (2007). Additive and classical drude polarizable force fields for linear and cyclic ethers. *J. Chem. Theory Comput.*, **3**, 1120–1133.

Wang, L., Du, W., Liu, D., and Dai, N. (2006). Lipase-catalyzed biodiesel production from soybean oil deodorizer distillate with absorbent present in *tert*-butanol system. *J. Mol. Cat. B: Enzym.*, **43**, 29–32.

Wang, S. S., Gray, C. G., Egelstaff, P. A., and Gubbins, K. E. (1973). Monte Carlo study of the pair correlation function for a liquid with non-central forces. *Chem. Phys. Lett.*, **21**, 123–126.

Wangikar, P. W., Graycar, T. P., Estell, D. A., Clark, D. S., and Dordick, J. S. (1993). Protein and solvent engineering of subtilisin BPN' in nearly anhydrous organic media. *J. Am. Chem. Soc.*, **115**, 12231–12237.

Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, **103**, 227–249.

Watanabe, K., Yoshida, T., and Ueji, S. (2004). The role of conformational flexibility of enzymes in the discrimination between amino acid and ester substrates for the subtilisin-catalyzed reaction in organic solvents. *Bioorg. Chem.*, **32**, 504–515.

Wedberg, R., Peters, G. H., and Abildskov, J. (2008). Total correlation function integrals and isothermal compressibilities from molecular simulations. *Fluid Phase Equilib.*, **273**, 1–10.

Wedberg, R., O'Connell, J. P., Peters, G. H., and Abildskov, J. (2010). Accurate Kirkwood-Buff integrals from molecular simulations. *Mol. Simul.*, **In press**.

Weerasinghe, S. and Smith, P. E. (2003). Kirkwood-Buff derived force field for mixtures of acetone and water. *J. Chem. Phys.*, **118**, 10663–10670.

Wescott, C. R. and Klibanov, A. M. (1993a). Predicting the solvent dependence of enzymatic substrate specificity using semiempirical thermodynamic calculations. *J. Am. Chem. Soc.*, **115**, 10362–10363.

Wescott, C. R. and Klibanov, A. M. (1993b). Solvent variation inverts substrate specificity of an enzyme. *J. Am. Chem. Soc.*, **115**, 1629–1631.

Wooley, R. J. and O'Connell, J. P. (1991). A database of fluctuation thermodynamic properties and molecular correlation function integrals for a variety of binary liquids. *Fluid Phase Equilib.*, **66**, 233–261.

Xu, Z., Affleck, R., Wangikar, P. W., Suzawa, V., Dordick, J. S., and Clark, D. S. (1994). Transition state stabilization of subtilisins in organic media. *Biotechnol. Bioeng.*, **43**, 515–520.

Yang, L., Dordick, J. S., and Garde, S. (2004). Hydration of enzyme in nonaqueous media is consistent with solvent dependence of its activity. *Biophys. J.*, **87**, 812–821.

Yennawar, N. H., Yennawar, H. P., and Farber, G. K. (1994). X-ray crystal-structure of γ-chymotrypsin in hexane. *Biochemistry*, **33**, 7326–7336.

Zaks, A. and Klibanov, A. M. (1984). Enymatic catalysis in organic media at 100 °C. *Science*, **224**, 1249–1251.

Zaks, A. and Klibanov, A. M. (1985). Enzyme-catalyzed processes in organic solvents. *Proc. Natl. Acad. Sci. USA*, **82**, 3192–3196.

Zaks, A. and Klibanov, A. M. (1988a). The effect of water on enzyme action in organic media. *J. Biol. Chem.*, **263**, 8017.

Zaks, A. and Klibanov, A. M. (1988b). Enzymatic catalysis in nonaqueous solvents. *J. Biol. Chem.*, **263**, 3194–3201.

Zheng, Y. and Ornstein, R. L. (1996a). A molecular dynamics and quantum mechanics analysis of the effect of DMSO on enzyme structure and dynamics: Subtilisin. *J. Am. Chem. Soc.*, **118**, 4175–4180.

Zheng, Y. and Ornstein, R. L. (1996b). Molecular dynamics of subtilisin carlsberg in aqueous and nonaqueous solutions. *Biopolymers*, **38**, 791–799.

Zheng, Y. and Ornstein, R. L. (1996c). A molecular dynamics study of the effect of carbon tetrachloride on enzyme structure and dynamics: Subtilisin. *Protein Eng.*, **9**, 485–492.

List of Abbreviations

BPTI	Bovine pancreatic trypsin inhibitor
CALB	*Candida antarctica* lipase B
CAMD	Computer-aided molecular design
COM	Center of mass
DCF	Direct correlation function
DCFI	Direct correlation function integral
DFT	Density functional theory
DMF	Dimethyl formamide
DMSO	Dimethyl sulphoxide
GEMC	Gibbs-ensemble Monte Carlo
GMD	Grand canonical ensemble molecular dynamics
EOS	Equation of state
FST	Fluctuation solution theory
HNC	Hypernetted chain
KB	Kirkwood-Buff
LJ	Lennard-Jones
MAG	Monoacylglycerol
MC	Monte Carlo
MD	Molecular dynamics
mM	Modified Margules
MTBE	Methyl t-butyl ether
OZ	Ornstein-Zernike
PY	Percus-Yevick
RDF	Radial distribution function
RML	*Rhizomucor miehei* lipase
RMSD	Root-mean square deviation
SASA	Solvent-accessible surface area
TCF	Total correlation function
TCFI	Total correlation function integral
TcTIM	*Trypanosoma cruzi* triosephosphate isomerase
TST	Transition state theory
UNIFAC	Universal Functional Activity Coefficient

This PhD-project was carried out at CAPEC, the Computer Aided Product-Process Engineering Center.
CAPEC is committed to research, to work in close collaboration with industry and to participate in educational activities. The research objectives of CAPEC are to develop computer-aided systems for product/process simulation, design, analysis and control/operation for chemical, petrochemical, pharmaceutical and biochemical industries. The dissemination of the research results of CAPEC is carried out in terms of computational tools, technology and application. Under computational tools, CAPEC is involved with mathematical models, numerical solvers, process/operation mathematical models, numerical solvers, process simulators, process/product synthesis/design toolbox, control toolbox, databases and many more. Under technology, CAPEC is involved with development of methodologies for synthesis/design of processes and products, analysis, control and operation of processes, strategies for modelling and simulation, solvent and chemical selection and design, pollution prevention and many more. Under application, CAPEC is actively involved with developing industrial case studies, tutorial case studies for education and training, technology transfer studies together with industrial companies, consulting and many more.

Further information about CAPEC can be found at www.capec.kt.dtu.dk.

Computer Aided Process Engineering Center
Department of Chemical and Biochemical Engineering
Technical University of Denmark
Søltofts Plads, Building 229
DK-2800 Kgs. Lyngby
Denmark

Phone: +45 4525 2800
Fax: +45 4525 4588
Web: www.capec.kt.dtu.dk

ISBN : 978-87-92481-45-0